

LA FILOSOFÍA POLÍTICA DE LA INTELIGENCIA ARTIFICIAL

UNA INTRODUCCIÓN

Mark Coeckelbergh



CÁTEDRA

Mark Coeckelbergh

*La filosofía política
de la inteligencia artificial*
Una introducción

Traducción de Lucas Álvarez Canga

CÁTEDRA
TEOREMA

Índice

AGRADECIMIENTOS

CAPÍTULO PRIMERO. Introducción. «Sospecho que el ordenador se equivocó»: Josef K. en el siglo XXI

Justificación, objetivos y enfoque de este libro

La estructura del libro y el resumen de sus capítulos

CAPÍTULO 2. Libertad manipulación por parte de la IA y esclavitud robótica

Introducción: declaraciones históricas de libertad y esclavitud contemporánea

IA, vigilancia y la aplicación de la ley: la eliminación de la libertad negativa

IA y la dirección del comportamiento humano: esquivando la autonomía humana

Amenazas a la autorrealización y a la emancipación: la explotación mediante el uso de la IA y el problema de los robots esclavos

¿Quién decide sobre la IA? Libertad como participación, IA en las elecciones y libertad de expresión

Otras nociones políticamente relevantes de la libertad y otros valores

CAPÍTULO 3. Igualdad y justicia: sesgo y discriminación por la IA

Introducción: sesgo y discriminación como eje central para plantear problemas relacionados con la igualdad y la justicia

¿Por qué está mal el sesgo (1)? Igualdad y justicia en la filosofía política liberal anglosajona estándar

¿Por qué está mal el sesgo (2)? Las teorías de la identidad y de clase
como crítica del pensamiento liberal universalista

Conclusión: la IA no es políticamente neutral

CAPÍTULO 4. Democracia: cámaras de eco y totalitarismo de las máquinas

Introducción: la IA como amenaza para la democracia

La IA como amenaza para la democracia, el conocimiento, la
deliberación y la política misma

Comenzando con Platón: democracia, conocimiento y experiencia

Más allá del gobierno de la mayoría y de la representación

Democracia deliberativa y participativa contra democracia agonística
y radical

Burbujas de información, cámaras de eco y populismo

Más problemas: manipulación, sustitución, responsabilidad y poder

La IA y los orígenes del totalitarismo: lecciones de Arendt

IA y totalitarismo

Arendt sobre los orígenes del totalitarismo y de la banalidad del mal

CAPÍTULO 5. El poder: la vigilancia y la (auto)disciplina por medio de los datos

Introducción: el poder como tema de la filosofía política

El poder y la IA: hacia un marco conceptual general

El marxismo: la IA como herramienta para el tecnocapitalismo

Foucault: cómo la IA nos somete y nos convierte en sujetos

Castigo y vigilancia

Conocimiento, poder y la creación y modelación de los sujetos y los
yoes

Tecnoperformances, poder e IA

Conclusión y preguntas pendientes

CAPÍTULO 6. ¿Qué sucede con los no humanos? Política medioambiental y poshumanismo

Introducción: más allá de una política antropocéntrica de la IA y la robótica

No solo los humanos cuentan políticamente: el estatus político de los animales (no humanos) y de la naturaleza

Implicaciones para la política de la IA y la robótica

El significado político del impacto de la IA sobre los no humanos y los entornos naturales

¿Estatus político para la propia IA?

Capítulo 7. Conclusión: tecnologías políticas

Qué he hecho en este libro y qué podemos concluir

Qué se necesita hacer a continuación: la cuestión respecto de las tecnologías políticas

BIBLIOGRAFÍA

CRÉDITOS

Agradecimientos

Me gustaría darle las gracias a mi editor, Mary Savigar, por su apoyo y por conducir el proyecto de este libro a un final feliz, a Justin Dyer por su cuidadosa edición, y a Zachary Storms por ayudarme con los aspectos organizativos relacionados con la presentación de este manuscrito. También quiero dar las gracias a los revisores anónimos por sus comentarios, que me ayudaron a pulir el manuscrito. Estoy especialmente agradecido a Eugenia Stamboliev por ayudarme en la búsqueda de bibliografía para este libro. Finalmente, quiero dar sinceramente las gracias a mi familia y a mis amigos, tanto próximos como distantes, por su apoyo durante estos dos difíciles años.

CAPÍTULO PRIMERO

Introducción. «Sospecho que el ordenador se equivocó»: Josef K. en el siglo XXI

Alguien ha tenido que haber estado contando mentiras sobre Josef K., pues fue arrestado una mañana sin haber hecho nada malo (Kafka, 2009, 5).

Esta es la primera línea de *El proceso* de Franz Kafka, libro publicado originalmente en 1925 y ampliamente considerado como una de las novelas más importantes del siglo xx. El protagonista de esta historia, Josef K., es arrestado y procesado, pero no sabe por qué. Al lector no se le dice nada sobre esta cuestión. Se presentan a continuación muchas exploraciones y encuentros que solo incrementan la oscuridad de todo el asunto y, tras un juicio injusto, Josef K. es ejecutado con un cuchillo de carnicero, «como un perro» (165). La historia se ha interpretado de muchas formas. En forma política se considera que es una muestra de cuán opresivas pueden ser las instituciones y que sus descripciones no reflejan solamente el incremento del poder de las burocracias modernas, sino que también prefiguran los horrores del régimen nazi que tuvieron lugar una década más tarde: la gente fue arrestada sin haber hecho nada malo y se la envió a campos de concentración, enfrentándose a diferentes formas de sufrimiento y, a menudo, a la muerte. Como señala Adorno: Kafka ofrece una «profecía del terror y de la tortura que se cumplió» (Adorno, 1983, 259).

Desafortunadamente, la historia de Kafka es aún relevante hoy día. No solo porque existan todavía burocracias opacas y regímenes opresivos que arrestan a personas sin justificación y, a veces, sin juicio, o porque (como ya señalaron Arendt [1943] y Agamben [1998]) los refugiados sufren a menudo un destino similar, sino también porque existe ahora una nueva forma en la que puede ocurrir todo esto, en la que ciertamente *ha* ocurrido,

incluso en las consideradas sociedades «avanzadas»: una forma que tiene que ver con la tecnología, en particular con la inteligencia artificial (IA).

En la tarde de un jueves de enero de 2020, Robert Julian-Borchak Williams recibió una llamada de teléfono en su oficina proveniente del Departamento de Policía de Detroit: se le solicitaba que se dirigiera a la comisaría de Policía para ser detenido. Puesto que no había hecho nada malo, no fue donde se le indicó. Una hora más tarde fue detenido en el jardín de la parte delantera de su casa, en presencia de su mujer y de sus hijos y, de acuerdo con el *New York Times*: «la policía no sabía por qué» (Hill, 2020). Más tarde, en la sala de interrogatorios, los policías le mostraron una imagen de una cámara de vigilancia en la que aparecía un hombre negro robando en una *boutique* de lujo y le preguntaron: «¿Es este usted?». El señor Williams, que era afroamericano, respondió: «No, este no soy yo. ¿Creéis que todos los negros somos iguales?». Fue liberado mucho más tarde, y al final el fiscal se disculpó.

¿Qué había ocurrido? El periodista del *New York Times* y los expertos que había consultado sospechaban que «su caso podría ser el primero conocido de un americano que había sido arrestado erróneamente basándose en un resultado erróneo de un algoritmo de reconocimiento facial». El sistema de reconocimiento facial, que utiliza IA en la forma de aprendizaje automático, tiene fallos y, muy probablemente, también esté sesgado: funciona mejor con hombres blancos que con otras comunidades. De este modo, el sistema crea falsos positivos, como en el caso del señor Williams y, combinado con un mal trabajo policial, tiene como resultado que se arreste a personas por crímenes que no cometieron. «Sospecho que el ordenador se equivocó», dijo uno de los agentes. En los Estados Unidos del siglo XXI, Josef K. es negro y se le acusa falsamente por medio de un algoritmo, sin mayor explicación.

La moraleja de la historia es que los ordenadores no solo cometen errores, errores que pueden conllevar serias consecuencias para personas particulares y sus familias. El uso de la IA también puede empeorar las injusticias y desigualdades sistémicas ya existentes y, en respuesta a casos como los del señor Williams, se podría alegar que todos los ciudadanos deberían tener el derecho a una explicación cuando se toman decisiones que

les conciernen. Además, esta es *una* de las muchas formas en las que la IA puede tener importancia e impacto político, intencionado a veces, pero a menudo sin que lo sea. Este caso particular plantea ciertas cuestiones relacionadas con el racismo y la (in)justicia: dos cuestiones oportunas. Pero hay mucho más que decir sobre las políticas de la IA y de las tecnologías relacionadas.

JUSTIFICACIÓN, OBJETIVOS Y ENFOQUE DE ESTE LIBRO

Mientras que las cuestiones *éticas* planteadas por la IA y las tecnologías relacionadas como la robótica y la automatización reciben actualmente una gran atención (Bartneck *et al.*, 2021; Boddington, 2017; Bostrom, 2014; Coeckelbergh, 2020; Dignum, 2019; Dubber, Pasquale y Das, 2020; Gunkel, 2018; Liao, 2020; Lin, Abney y Jenkins, 2017; Nyholm, 2020; Wallach y Allen, 2009), existen muy pocos trabajos que aborden el tema desde una perspectiva *político-filosófica*. Esta situación es lamentable, ya que la cuestión se presta perfectamente a una investigación de este tipo y no se beneficia de los valiosos recursos intelectuales de la tradición político-filosófica. Por su parte, la mayoría de los *filósofos políticos* no han abordado para nada el tema de la política de la IA (las excepciones son Benjamin, 2019a; Binns, 2018; Eubanks, 2018; Zimmermann, Di Rosa y Kim, 2020), a pesar de que, en general, hay un creciente interés en el tema, por ejemplo en cómo se usan los algoritmos y el *big data* de maneras que refuerzan tanto el racismo como las diferentes formas de desigualdad e injusticia (por ejemplo, Bartoletti, 2020; Criado Perez, 2019; Noble, 2018; O’Niel, 2016) y que extraen y consumen recursos planetarios (Crawford, 2021).

Además, mientras que en el *contexto político* actual hay una gran atención pública dirigida hacia cuestiones como la libertad, la esclavitud, el racismo, el colonialismo, la democracia, la competencia técnica, el poder y el clima, a menudo se discuten estos temas de una forma que parece que tuvieran poco que ver con la tecnología, y viceversa. La IA y la robótica se consideran temas técnicos, y si se crea una relación con la política, la tecnología se considera como una herramienta utilizada para la

manipulación política o la vigilancia. Normalmente quedan sin abordar los efectos no deseados. Por otro lado, *los desarrolladores y los científicos* que trabajan en los campos de la IA, la ciencia de datos y la robótica están a menudo dispuestos a tener en cuenta en su trabajo cuestiones éticas, pero no son conscientes de los complejos problemas políticos y sociales con los que están conectadas estas cuestiones, por no mencionar las sofisticadas discusiones político-filosóficas que pueden producirse sobre cómo estructurar y abordar estos problemas. Además, como la mayoría de la gente no está familiarizada con el pensamiento sistemático sobre la tecnología y la sociedad, tienden a dar por sentado el punto de vista de que la tecnología misma es neutral y que todo depende de los humanos que la desarrollan y la usan.

Cuestionar esta concepción naíf de la tecnología es la especialidad de la *filosofía de la tecnología*, que ha desarrollado, en su forma contemporánea, una comprensión no instrumental de la tecnología: la tecnología no es solamente un medio para alcanzar un fin, sino que también moldea ese fin (para tener una perspectiva general de algunas teorías, ver Coeckelbergh, 2019a). Sin embargo, en el momento en el que se utilizan marcos filosóficos y fundamentos conceptuales para la evaluación normativa de la tecnología, los filósofos de la tecnología corren normalmente hacia la ética (por ejemplo, Gunkel, 2014; Vallor, 2016). La filosofía política es ampliamente ignorada. Solo algunos filósofos establecen esta conexión: por ejemplo, en los años 80 y 90, Winner (1986) y Feenberg (1999) y, actualmente, Sattarov (2019) y Sætra (2020). Se necesita más investigación sobre la conexión entre la filosofía de la tecnología y la filosofía política.

Existe aquí una brecha académica, pero también una necesidad social. Si queremos abordar algunos de los problemas más acuciantes a nivel global y local del siglo XXI, como el cambio climático, las desigualdades globales, el envejecimiento, las nuevas formas de exclusión, la guerra, el autoritarismo, las epidemias y las pandemias, etc., todos ellos no solo políticamente relevantes, sino también relacionados con la tecnología de diferentes maneras, es importante crear un diálogo entre la reflexión sobre política y la reflexión sobre tecnología.

Este libro cubre estas brechas y responde a esa justificación mediante:

- La conexión entre cuestiones normativas sobre la IA y la robótica con discusiones clave de la filosofía política, usando tanto la historia de la filosofía política como trabajos más recientes.
- El abordaje de temas controvertidos que se encuentran en el centro de la actual atención política, pero relacionándolos ahora con cuestiones que competen a la IA y a la robótica.
- La presentación de cómo esto no es solo un ejercicio de filosofía política aplicada, sino que también conduce a intuiciones interesantes acerca de la, a menudo, oculta y más profunda dimensión política de estas tecnologías contemporáneas.
- La exposición de cómo las tecnologías de la IA y la robótica tienen efectos políticos intencionados y no intencionados, que pueden ser discutidos con efectos provechosos utilizando la filosofía política.
- La aportación, por consiguiente, de contribuciones originales tanto a la filosofía de la tecnología como a la filosofía política aplicada.

Así pues, este libro utiliza la filosofía política, codo con codo con la filosofía de la tecnología y la ética, con los objetivos de (1) comprender mejor las cuestiones normativas planteadas por la IA y la robótica y (2) arrojar luz sobre cuestiones políticas acuciantes y sobre la forma en que están imbricadas con el uso de estas nuevas tecnologías. Uso el término «imbricadas» para expresar la conexión cercana que existe entre las cuestiones políticas y las cuestiones que atañen a la IA. La idea es que la última de las cuestiones *ya* es política. El concepto rector de este libro es que la IA no es simplemente un asunto técnico o lo es solamente sobre inteligencia: no es neutral en términos de política y poder. La IA es *política de cabo a rabo*. Mostraré y discutiré en cada capítulo esa dimensión política de la IA.

En lugar de presentar una discusión sobre la política de la IA en general, plantearé este tema general acercándome a temas específicos que aparecen en la filosofía política contemporánea. Cada capítulo se centrará en un conjunto de temas filosófico-políticos concretos: la libertad, la manipulación, la explotación y la esclavitud; la igualdad, la justicia, el racismo, el sexismo y otras formas de sesgo y discriminación; la

democracia, la competencia técnica, la participación y el totalitarismo; el poder, el castigo, la vigilancia y la autoconstitución; los animales, el medioambiente y el cambio climático en relación con el poshumanismo y el transhumanismo. Cada tema se discutirá a la luz de los efectos, tanto los que se intentan conseguir como los que no, de la IA, de la ciencia de datos y de las tecnologías relacionadas como la robótica.

Como percibirá el lector, la división en temas y conceptos es, hasta cierto punto, artificial. Quedará patente que hay muchas formas en las que los conceptos, y, por ende, los temas y capítulos, se interrelacionan e interactúan. Por ejemplo, el principio de libertad puede estar en tensión con el principio de igualdad, y es imposible hablar de democracia e IA sin hablar sobre el poder. Algunas de estas conexiones se harán explícitas a lo largo del libro y otras corren a cuenta del lector, pero todos los capítulos muestran cómo la IA impacta en estas cuestiones políticas clave y cómo la IA es política.

Sin embargo, este libro no trata solo sobre la IA, sino también sobre el propio pensamiento filosófico-político. Estas discusiones sobre la política de la IA no serán solamente ejercicios de filosofía aplicada (más en concreto, filosofía política aplicada), sino que también retroalimentarán a los propios conceptos filosófico-políticos. Muestran cómo las nuevas tecnologías ponen en tela de juicio las nociones mismas de libertad, igualdad, democracia, poder, y así sucesivamente. ¿Qué significan estos principios políticos y conceptos filosófico-políticos en la edad de la IA y la robótica?

LA ESTRUCTURA DEL LIBRO Y EL RESUMEN DE SUS CAPÍTULOS

El libro está organizado en siete capítulos.

En el capítulo 2 planteo algunas preguntas relacionadas con el principio político de libertad. ¿Qué significa la libertad cuando la IA ofrece nuevas formas de tomar, manipular e influir en nuestras decisiones? ¿Cuán libres somos cuando realizamos trabajos digitales para grandes y poderosas empresas? ¿Y conlleva la sustitución de trabajadores por robots la continuación del pensamiento esclavista? El capítulo está estructurado de

acuerdo con las diferentes concepciones de la libertad. Discute las posibilidades ofrecidas por la toma algorítmica de decisiones y su capacidad de influencia, conectándolas con discusiones sobre la libertad en la filosofía política (libertad negativa y positiva) y la tesis del *nudge*¹ (empujón). Señala cómo se nos puede arrebatar la libertad negativa sobre la base de la recomendación de una IA, cuestiona cómo es realmente el *nudge* libertario a través de la IA y plantea preguntas críticas basadas en Hegel y Marx, mostrando cómo el significado y el uso de robots amenaza con mantenernos conectados a una historia y un presente de esclavitud y explotación capitalista. El capítulo finaliza con una discusión sobre la IA y la libertad entendida como participación política y libertad de expresión, que continúa en el capítulo 4 sobre la democracia.

El capítulo 3 pregunta: ¿cuáles son los efectos políticos de la IA y de la robótica (normalmente no intencionados) en términos de igualdad y justicia? ¿Incrementan las desigualdades en la sociedad la automatización y la digitalización que permite la robótica? ¿La toma de decisiones automatizada por parte de la IA conduce a una discriminación injusta, sexismo y racismo, como Benjamin (2019a), Noble (2018) y Criado Perez (2019) han argumentado, y, si es así, por qué? ¿Es problemática la sexualización de los robots, y cómo? ¿Cuál es el significado de justicia y equidad utilizado en estas discusiones? Este capítulo sitúa el debate sobre la automatización y la discriminación por parte de la IA y de la robótica en el contexto de las discusiones filosófico-políticas clásicas sobre la (des)igualdad y la (in)justicia como equidad en la tradición filosófica liberal (por ejemplo, Rawls, Hayek), pero también conecta con el marxismo, el feminismo crítico y el pensamiento antirracista y anticolonial. Plantea también preguntas relacionadas con la tensión entre las concepciones de la justicia universal frente a la justicia basada en la identidad de grupo y la discriminación positiva, y discute algunas cuestiones sobre temas relativos a la justicia intergeneracional y la justicia global. El capítulo concluye con la tesis de que los algoritmos de la IA no son nunca políticamente neutrales.

En el capítulo 4 discuto los impactos de la IA sobre la democracia. La IA puede usarse para manipular a los votantes y las elecciones. ¿Destruye la democracia la vigilancia mediante IA? ¿Sirve al capitalismo, como ha

defendido Zuboff (2019)? Y ¿nos encontramos de camino hacia un tipo de «fascismo y colonialismo de datos»? En cualquier caso, ¿qué queremos decir con «democracia»? Este capítulo sitúa las discusiones sobre la democracia y la IA en el contexto de la teoría de la democracia, de las discusiones sobre el papel de la competencia técnica en la política, y se ocupa de las condiciones para el totalitarismo. En primer lugar, muestra que, mientras que es sencillo ver cómo la IA puede amenazar la democracia, es mucho más difícil hacer explícito qué tipo de democracia queremos y cuál es y debería ser el papel de la tecnología en la democracia. El capítulo bosqueja las tensiones entre las concepciones platónico-tecnocráticas de la política y los ideales de la democracia participativa y deliberativa (Dewey y Habermas), que, a su vez, tienen sus propios críticos (Mouffe y Rancière). Conecta esta discusión con cuestiones como las burbujas de información, las cámaras de eco y el populismo impulsado por la IA. En segundo lugar, el capítulo argumenta que el problema del totalitarismo a través de la tecnología nos conduce a problemas más profundos y duraderos de la sociedad moderna, como la soledad (Arendt) y la falta de confianza. Las discusiones éticas, siempre y cuando se centren en el daño a los individuos, olvidan esta dimensión social e histórica más amplia. El capítulo termina señalando el peligro de lo que Arendt (2006) denominó «la banalidad del mal» cuando la IA se utiliza como una herramienta para la manipulación de las grandes corporaciones y para la gestión burocrática de las personas.

El capítulo 5 discute la IA y el poder. ¿Cómo se puede usar la IA para sancionar la observancia de las leyes y autodisciplinarse? ¿Cómo afecta al conocimiento, cambia y moldea las relaciones de poder existentes entre humanos y máquinas, pero también entre los propios humanos e, incluso, dentro de los humanos? ¿Quién se beneficia de ello? Para plantear estas cuestiones, el capítulo conecta de nuevo con las discusiones sobre la democracia, la vigilancia y el capitalismo de la vigilancia, pero también introduce la compleja visión del poder de Foucault que subraya los micromecanismos del poder al nivel de las instituciones, de las relaciones humanas y de los cuerpos. En primer lugar, el capítulo desarrolla un marco conceptual con el que reflexionar sobre las relaciones entre el poder y la IA.

A continuación, echa mano de tres teorías del poder para elaborar alguna de estas relaciones: el marxismo y la teoría crítica, Foucault y Butler y el enfoque orientado hacia la *performatividad*. Esto me permite arrojar luz sobre los atractivos y las manipulaciones de y por la IA, la explotación y la autoexplotación que produce en su contexto capitalista, y la historia de la ciencia de datos en términos de señalar, clasificar y vigilar a las personas. Pero también indica formas en las que la IA puede empoderar a las personas y, a través de las redes sociales, desempeñar un papel en la constitución del yo y de la subjetividad. Además, se argumenta que, viendo lo que aquí logran los humanos y la IA en términos de *tecno-performatividad*, podemos señalar el papel cada vez más director y más-que-instrumental que desempeña la tecnología a la hora de organizar las formas en que nos movemos, actuamos y sentimos. Muestro que estos ejercicios de (tecno)poder siempre tienen una dimensión activa y social que implica tanto a la IA como a los humanos.

En el capítulo 6 introduzco algunas cuestiones relacionadas con los no humanos. Como la mayoría de la ética de la IA, las discusiones políticas clásicas son humanocéntricas, pero esto puede ser, y lo ha sido de hecho, cuestionado al menos de dos formas. Primero, ¿cuentan políticamente solo los humanos? ¿Cuáles son las consecuencias de la IA para los no humanos? Y, ¿es la IA una amenaza o una oportunidad para lidiar con el cambio climático? ¿O es ambas cosas? Segundo, ¿pueden tener los sistemas de IA y los propios robots un estatus político de, por ejemplo, ciudadanía? Los poshumanistas cuestionan la visión tradicionalmente antropocéntrica de la política. Además, los transhumanistas han argumentado que los humanos serán reemplazados por agentes artificiales superinteligentes. ¿Cuáles son las implicaciones políticas si un ser superinteligente se pone al mando? ¿Es el fin de la libertad humana, la justicia y la democracia? Utilizando recursos de los derechos animales y de la teoría medioambiental (Singer, Cochrane, Garner, Rowlands, Donaldson y Kimlicka, Callicott, Rolston, Leopold, etc.), del poshumanismo (Haraway, Wolfe, Braidotti, Massumi, Latour, etc.), de la ética de la IA y de la robótica (Floridi, Bostrom, Gunkel, Coeckelbergh, etc.) y del transhumanismo (Bostrom, Kurzweil, Moravec, Hughes, etc.), este capítulo explora las concepciones de la política de la IA

que van más allá de lo humano. Argumento que tal política necesitaría el replanteamiento de nociones tales como libertad, justicia y democracia para incluir a los no humanos, y que plantearía nuevas cuestiones para la IA y la robótica. El capítulo finaliza con la afirmación de que la política no antropocéntrica de la IA remodela ambos términos de la relación humano-IA: los humanos no están solamente *des-empoderados* y *empoderados* por la IA, sino que también otorgan a la IA su poder.

El capítulo final resume el libro y concluye que (1) las cuestiones de las que nos preocupamos actualmente en las discusiones políticas y sociales como, por ejemplo, la libertad, el racismo, la justicia y la democracia adquieren una nueva urgencia y relevancia a la luz de desarrollos tecnológicos tales como la IA y la robótica; y que (2) conceptualizar la política de la IA y la robótica no es asunto de aplicar simplemente las nociones existentes que provienen de la filosofía y la teoría políticas, sino que nos invitan a cuestionar las nociones mismas (libertad, igualdad, justicia, democracia, etc.) y a plantear preguntas interesantes sobre la naturaleza y el futuro de la política, así como sobre nosotros mismos como humanos. El capítulo también defiende que, dada la estrecha imbricación de la tecnología con los cambios y transformaciones sociales, medioambientales y psicológico-existenciales, la filosofía política del siglo XXI no puede seguir evitando lo que Heidegger (1977) denominó «la pregunta sobre la técnica». El capítulo también perfila algunos pasos posteriores que se deberían dar en este campo. Necesitamos más filósofos que trabajen en esta área y más investigación sobre el nexo de la filosofía política/filosofía de la tecnología que, esperemos, conduzca a un ulterior «pensar juntos» (*zusammendenken*) sobre la política y la tecnología. También necesitamos reflexionar más sobre cómo hacer la política de la IA más participativa, pública, democrática, inclusiva y sensible a los contextos globales y diferencias culturales. El libro termina con esta pregunta: ¿qué *tecnologías políticas* necesitamos para moldear ese futuro?

¹ El término *nudge* originalmente hace referencia a la acción de empujar suavemente algo o a alguien (preferentemente con el codo) para llamar su atención. Aquí se usa como elemento que, de alguna manera sutil, inclina a un agente a realizar una acción. Por ejemplo, la colocación en determinado

lugar de ciertos productos en un supermercado «empuja» (i. e. «*nudges*») a los clientes a comprarlos.
[N. del T.]

CAPÍTULO 2

Libertad: manipulación por parte de la IA y esclavitud robótica

INTRODUCCIÓN: DECLARACIONES HISTÓRICAS DE LIBERTAD Y ESCLAVITUD CONTEMPORÁNEA

La libertad (tanto positiva como negativa)² es considerada como uno de los principios políticos más importantes de las democracias liberales, cuyas constituciones buscan proteger las libertades básicas de los ciudadanos. Por ejemplo, la Primera Enmienda de la Constitución de los Estados Unidos, adoptada en 1791 como parte de la Carta de Derechos, protege las libertades individuales como la libertad de religión, libertad de expresión y el derecho de reunión. La Constitución alemana o ley fundamental (*Grundgesetz*), adoptada en 1949, afirma que la libertad de la persona es inviolable (artículo 2). Históricamente, la Declaración de los Derechos del Hombre y del Ciudadano de 1789 ha sido muy influyente. Hunde sus raíces en el pensamiento ilustrado (Rousseau y Montesquieu) y se desarrolló durante la Revolución francesa con la ayuda de Thomas Jefferson, uno de los padres fundadores de los Estados Unidos y autor principal de la Declaración de Independencia de los Estados Unidos de 1776, que ya proclamaba en su preámbulo que «todos los hombres son creados iguales» y que tienen «derechos inalienables», incluyendo los de «la vida, la libertad y la búsqueda de la felicidad». El artículo I de la Declaración francesa establece que «los hombres nacen y permanecen libres e iguales en derechos». Aunque esta Declaración aún excluía a las mujeres y no prohibía la esclavitud, fue parte de una serie de declaraciones de derechos y libertades civiles que comenzó en 1215 con la carta magna (*Magna Carta Libertatum* o carta suprema de derechos) y finalizó con la Declaración Universal de los Derechos Humanos (DUDH), adoptada por la Asamblea

General de las Naciones Unidas en diciembre de 1948, que establece que «todos los seres humanos nacen libres e iguales en dignidad y derechos» (artículo 1) y que «nadie estará sometido a esclavitud ni a servidumbre» (artículo 4) (UN, 1948).

Sin embargo, en muchos países del mundo la gente aún sufre por, y protesta contra, regímenes opresivos y autoritarios que amenazan o violan su libertad. A menudo, las protestas tienen consecuencias letales. Considérese, por ejemplo, cómo se trata a la oposición política en las contemporáneas Turquía, Bielorrusia, Rusia, China y Myanmar. Y, mientras que la esclavitud es ilegal, nuevas formas de esclavitud siguen vigentes en la actualidad. La Organización Internacional del Trabajo estima que hay globalmente más de 40 millones de personas que están sometidas a algún tipo de trabajo o explotación sexual forzados, por ejemplo en el trabajo doméstico o en la industria sexual (OIT, 2017). Esto ocurre dentro de los países y a través de la trata de personas. Están especialmente afectados mujeres y niños. Sucede en Corea del Norte, Eritrea, Burundi, República Centroafricana, Afganistán, Pakistán e Irán, pero también persiste en países como los Estados Unidos y Reino Unido. De acuerdo con el *Global Slavery Index* (Índice Global de Esclavitud), en 2018 se estimaba que había 403000 personas trabajando bajo condiciones de trabajo forzoso en los Estados Unidos (Walk Free Foundation, 2018, 180). Los países occidentales también importan bienes y servicios que tienen el riesgo de haber estado implicados en el trabajo esclavo moderno en su lugar de producción.

Pero ¿qué significa exactamente la libertad y qué significa la libertad política a la luz de los desarrollos de la IA y de la robótica? Para responder a estas preguntas, analizaremos diferentes amenazas a la libertad o, más bien, amenazas a *diferentes tipos de libertades*. Examinemos, pues, algunas concepciones clave de la libertad desarrolladas por filósofos políticos: libertad negativa, libertad como autonomía, libertad como autorrealización y emancipación, libertad como participación política y libertad de expresión.

IA, VIGILANCIA Y LA APLICACIÓN DE LA LEY: LA ELIMINACIÓN DE LA LIBERTAD NEGATIVA

Según hemos visto en la introducción, la IA puede emplearse para fomentar la aplicación de la ley. También puede usarse en la vigilancia de fronteras y en la seguridad aeroportuaria. Alrededor del mundo, la tecnología de reconocimiento facial y otras tecnologías biométricas como escáneres de huellas dactilares y de iris están siendo empleadas en aeropuertos y otros lugares de cruce de fronteras. Estas tecnologías corren el riesgo de producir sesgos y discriminaciones (ver el siguiente capítulo) y de amenazar la privacidad (UNICRI e INTERPOL, 2019). Todo ello puede derivar en múltiples intervenciones que infringen la *libertad* de la persona, incluyendo el arresto y el encarcelamiento. Si la tecnología de IA comete un error (por ejemplo, categorizando erróneamente a una persona al no reconocer correctamente su cara), los individuos pueden ser arrestados equivocadamente, se les puede denegar el asilo, ser acusados públicamente, etc., etc. Un «pequeño» margen de error puede afectar a miles de viajeros (Israel, 2020). De forma similar, la llamada *policía predictiva*, que utiliza aprendizaje automático para «predecir» delitos, puede conducirnos a decisiones judiciales que priven de libertad de forma injustificada, además de dar lugar a discriminación (de nuevo). Dicho de manera más general: todo esto puede conducirnos a situaciones «kafkianas», procesos opacos de tomas de decisión y decisiones arbitrarias, injustificadas e inexplicadas, que afectan enormemente a las vidas de los acusados y amenazan el Estado de derecho (Radavoi, 2020, 111-113; ver también Hildebrandt, 2015).

El tipo de libertad que está en riesgo aquí es el que los filósofos políticos denominan «libertad negativa». Como es bien sabido, Berlin definió la libertad negativa como libertad de interferencia. Se ocupa de la pregunta siguiente: ¿cuál es el área dentro de la cual al sujeto (una persona o un grupo de personas) se le deja (o debería dejársele) hacer o ser lo que quiera hacer o ser, sin la interferencia de otras personas?» (Berlin, 1997, 194). Así pues, la libertad negativa es la ausencia de interferencia, coerción u obstrucción por parte de otros o del Estado. Este es el tipo de libertad que está en juego cuando se usa la IA para identificar a personas que

constituyen una amenaza a la seguridad, a quienes se considera que no tienen derecho a emigrar o al asilo, o a quienes han cometido un delito. La libertad que está amenazada es la libertad de no interferencia.

A la luz de las tecnologías de la vigilancia, se podría extender esta concepción de la libertad a la libertad de no estar en *riesgo* de interferencia. Esta libertad negativa está en juego cuando se utiliza la tecnología de la IA para vigilancia, para mantener a ciertas personas en un estado de esclavitud o explotación. La tecnología crea cadenas invisibles y ojos no humanos que están vigilando constantemente. La cámara o el robot están siempre ahí. Como se ha observado a menudo, esta situación es similar a la que primero Bentham, y luego Foucault, denominaron el panóptico: los prisioneros estaban siempre vigilados, pero no podían ver a sus vigilantes (ver también el capítulo 5 sobre el poder). La reclusión física o la supervisión directa, como en las formas anteriores de encarcelamiento o esclavitud, deja de ser necesaria: es suficiente con que la tecnología esté ahí para monitorizar a las personas. Hablando técnicamente, ni siquiera tiene que estar funcionando. Compárese esto con las cámaras que controlan el tráfico: tanto si funcionan realmente como si no, influyen (en concreto, *adiestran*) en el comportamiento humano. Al saber que estás siendo controlado todo el tiempo, o que puedes estar siendo visto, eso es ya suficiente para introducir disciplina. Es suficiente que haya un riesgo de interferencia: crea el temor de que se nos despoje de nuestra libertad negativa. Puede utilizarse en prisiones y campos de concentración, pero también funciona en situaciones laborales para monitorizar el desempeño de los empleados. A menudo, la vigilancia se encuentra oculta. No vemos los algoritmos, los datos, ni aquellos que los usan. Debido a este aspecto oculto, Bloom (2019) habla de una forma un tanto engañosa sobre «poder virtual». Pero el poder es real.

La vigilancia por parte de la IA no se emplea solamente para fomentar la aplicación de la ley por los Gobiernos, o en entornos empresariales y contextos laborales, también se emplea en la esfera privada. Por ejemplo, en las redes sociales no hay solamente una vigilancia «vertical» (por parte del Estado y la empresa de la red social), sino también vigilancia por pares u «horizontal»: usuarios de la red social que se vigilan unos a otros, mediados por algoritmos. Y existe también la *sousveillance* (Mann, Nolan y Wellman,

2002): gente que usa dispositivos portátiles para grabar lo que está ocurriendo. Esto es problemático por diversos motivos, pero una razón es que amenaza la libertad. Esto podría significar aquí la libertad negativa de tener privacidad, entendida como libertad de interferencia en la esfera personal. La privacidad es considerada normalmente como un derecho básico en la sociedad liberal, esto es, libre. Pero puede encontrarse en peligro en una sociedad en la que se nos pide adoptar la cultura del compartir. Como Véliz (2020) señala:

El liberalismo pide que nada más esté sujeto al escrutinio público que lo que sea necesario para proteger a los individuos y para cultivar una vida colectiva física y moralmente saludable. Una cultura de la exposición requiere que todo sea compartido y esté sujeto al escrutinio público (110).

La transparencia completa amenaza entonces a las sociedades liberales, y los gigantes tecnológicos desempeñan un papel importante en este punto. Al usar redes sociales, creamos voluntariamente dossiers digitales sobre nosotros mismos con todo tipo de información personal y detallada que compartimos voluntariamente, sin ningún Gobierno de tipo Gran Hermano que nos fuerce a entregar esa información o teniendo que realizar el meticuloso trabajo de adquirirla de forma encubierta. Más bien, las compañías tecnológicas obtienen los datos de forma transparente y descarada. Plataformas como Facebook son el sueño húmedo de un régimen autoritario, pero también el de un capitalista. Las personas crean dossiers y se rastrean *a sí mismas*, por ejemplo con propósitos sociales (encuentros), pero también monitorizan su salud.

Además, tal información puede, y ha sido, de hecho, utilizada contra personas para fomentar el cumplimiento de la ley. Por ejemplo, basándose en el análisis de datos de su dispositivo Fitbit, un monitorizador de la salud, la policía de los Estados Unidos acusó a una mujer de haber hecho un informe falso sobre una violación (Kleeman, 2015). También se usaron en los Estados Unidos datos de Fitbit en un caso de asesinato (BBC, 2018). Los datos de las páginas de redes sociales y de los teléfonos pueden usarse como policía predictiva, lo que puede conllevar consecuencias para la libertad personal. Incluso si no existe una amenaza a la libertad de interferencia, el problema reside también en el nivel social e impacta sobre

diferentes tipos de libertades, tales como la libertad como autonomía (ver la siguiente sección). Como afirma Solove (2004): «Es un problema que está imbricado en el tipo de sociedad en que nos estamos convirtiendo, en la forma en que pensamos, en nuestro lugar en un orden social más extenso y en nuestra capacidad para ejercer un control significativo sobre nuestras vidas» (35).

Dicho esto, cuando se trata de las amenazas a la libertad negativa mediante la tecnología, la cuestión puede convertirse en algo de carácter muy físico. Los robots se pueden usar para contener físicamente a las personas, por ejemplo por motivos de seguridad, o para hacer que se cumpla la ley, pero también por el «propio bien de las personas» y su seguridad. Considérese la situación de un niño pequeño o de una persona mayor con discapacidad cognitiva que se arriesga a cruzar una calle peligrosa sin mirar o que corre el riesgo de caer por una ventana: en tales casos, se puede utilizar una máquina para contener a esa persona evitando, por ejemplo, que esa persona abandone su habitación o su casa. Esta es una forma paternalista (más en la siguiente sección) que restringe la libertad negativa mediante una vigilancia seguida de una forma física de interferencia. Sharkey y Sharkey (2012) ven incluso en el uso de robots para restringir las actividades de las personas mayores «una pendiente resbaladiza hacia una robótica autoritaria». Un escenario así, en el que se monitorice y se contenga a seres humanos usando tecnología de IA y robótica, parece más realista que el distante escenario de ciencia ficción de una IA superinteligente que tome el poder (que podría también conllevar la pérdida de libertad).

Quienquiera que use la IA o la robótica para restringir las libertades negativas de las personas tiene que justificar por qué es necesario violar un tipo tan básico de libertad. Como Mill (1963) argumentó a mediados del siglo XIX, cuando se trata de coerción, la carga de la prueba debería residir en aquellos que abogan por una restricción o prohibición, no en las personas que defienden sus libertades negativas. En el caso de la violación de la privacidad, de hacer que se cumpla la ley o de restricciones paternalistas del movimiento, la carga de la prueba reside en quien lleva a cabo la restricción. Este debe mostrar que existe un riesgo o daño considerable

(Mill), o que existe otro principio (por ejemplo, la justicia) que es más importante que la libertad (en general o en el caso particular). Y justificar tales usos e intervenciones se hace incluso más difícil cuando la tecnología comete errores (el caso del falso positivo de la introducción) o cuando la tecnología misma es la que causa el daño. Por ejemplo, el reconocimiento facial puede conducir a arrestos y encarcelamientos injustificados, o un robot puede causar daños cuando detiene a alguien. Además, más allá del marco utilitarista y, de forma más general, consecuencialista, se pueden subrayar los derechos a la libertad desde un punto de vista deontológico, por ejemplo los derechos a la libertad plasmados en las declaraciones nacionales e internacionales.

Sin embargo, al considerar estos casos en los que la tecnología tiene efectos dañinos (no intencionales), queda claro que hay más cosas en juego que simplemente la libertad. Existen tensiones y negociaciones entre la libertad y otros principios y valores políticos. La libertad negativa es muy importante, pero puede haber también otros principios políticos y éticos que sean también muy importantes y que desempeñen (o deberían hacerlo) algún papel en un caso particular. No siempre está claro cuál es el principio que debería prevalecer. Por ejemplo, mientras que puede estar muy claro que está justificado restringir la libertad de un niño pequeño para evitar un daño concreto (por ejemplo, caerse por una ventana), no queda tan claro que tales restricciones a la libertad estén justificadas en el caso de una persona anciana con demencia o en el caso de alguien que se considera que está viviendo «ilegalmente» en un país concreto. Y, ¿está justificado restringir la libertad negativa de una persona (por ejemplo, mediante el encarcelamiento) para proteger la libertad negativa y otros derechos políticos de otras personas?

La aplicación del principio del daño de Mill también es bastante complicada. ¿Qué constituye exactamente daño en un caso particular, quién define qué daño se le hace a quién y qué daño es más importante? Y, en cualquier caso, ¿qué cuenta como una restricción de la libertad negativa? Considérese, por ejemplo, la obligación de llevar puesta una mascarilla en lugares concretos durante la pandemia del COVID-19, que ha conducido a controversias sobre precisamente estas cuestiones: ¿quién necesita más

protección del (riesgo de) daño y está llevando una mascarilla dejándose arrebatar la libertad negativa? Estas cuestiones son también relevantes en el caso del uso de la IA. Por ejemplo, incluso si una concreta tecnología de IA funcionase sin errores, ¿el procedimiento de escaneo de seguridad de un aeropuerto, que implica el escaneo y reconocimiento facial, es en sí mismo una violación de mi libertad con la que no se debería interferir? ¿Es el cacheo una violación, y si es así, es una infracción mayor que el escaneo? ¿El error en el reconocimiento facial constituye un daño en sí mismo o depende de las potenciales acciones dañinas del personal de seguridad? Y si todo esto estuviera justificado apelando al riesgo del terrorismo, ¿justifica este riesgo con pequeña probabilidad (pero de alto impacto) las medidas que interfieren con mi libertad negativa cuando cruzo una frontera y me expongo a nuevos riesgos que crea la tecnología, incluyendo el riesgo de que se me arrebatase mi libertad negativa debido a un error tecnológico?

IA Y LA DIRECCIÓN DEL COMPORTAMIENTO HUMANO: ESQUIVANDO LA AUTONOMÍA HUMANA

Pero, si estas cuestiones afectan a la libertad negativa, ¿qué es la libertad positiva? Hay varias formas de definir la libertad positiva, pero un significado central definido por Berlin tiene que ver con la autonomía o el autogobierno. La cuestión aquí es si tu elección es realmente tu elección, más bien que la de otra persona. Berlin (1997) escribe:

El sentido «positivo» de la palabra «libertad» deriva del deseo, por parte del individuo, de ser su propio dueño. Deseo que mi vida y mis decisiones dependan de mí mismo, no de fuerzas externas de cualquier tipo. [...] Deseo, sobre todo, ser consciente de mí mismo como ser pensante, deseante y activo, que carga con la responsabilidad de sus elecciones y es capaz de explicarlas haciendo referencia a sus propias ideas y propósitos. (203)

Este tipo de libertad está en contraste no con la interferencia en el sentido de encarcelamiento u obstrucción, sino más bien con el paternalismo: alguna otra instancia decide lo que es mejor para ti. Berlin argumenta que los gobernantes autoritarios distinguen entre un yo superior y un yo inferior, afirman conocer lo que realmente quiere tu yo superior y, a continuación, oprimen a la gente en nombre de ese yo superior. Este tipo de

libertad no trata de la ausencia de una restricción externa o de un castigo físico: es una interferencia en tu psicología del deseo y de la elección.

¿Qué tiene esto que ver con la IA? Para entenderlo, considérese la posibilidad del *nudge*: cambiar el entorno de la elección con el objetivo de alterar el comportamiento de las personas. El concepto de *nudge* explota la psicología de la toma de decisiones humana, en particular los sesgos en la toma de tales decisiones. Thaler y Sunsteins (2009) proponen el *nudge* como una solución al problema de que no se puede confiar en la gente a la hora de tomar decisiones racionales y usan, en su lugar, la heurística y los sesgos para decidir. Argumentan que deberíamos influir en la toma de decisiones de las personas cambiando su entorno de elección en una dirección deseable. En lugar de forzar a la gente, alteramos su «estructura de elección» (6). Por ejemplo, en lugar de prohibir la comida basura, se ofrece la fruta a la altura de los ojos y en un lugar destacado del supermercado. Esta intervención opera entonces subcientemente; apunta a la parte del cerebro que compartimos con los lagartos (20). Ahora bien, la IA podría ser, y ha sido, usada para este tipo de *nudge*, por ejemplo cuando Amazon hace recomendaciones de productos que afirma que yo debería querer comprar. De forma similar, Spotify parece afirmar que me conoce mejor de lo que yo mismo me conozco cuando me recomienda una música concreta. Estos sistemas de recomendación actúan como *nudges* en el sentido de que no limitan mi elección de libros o música, pero influyen en mi compra, mi lectura y mi comportamiento como oyente en la dirección sugerida por el algoritmo. Y esta misma tecnología se puede utilizar o puede estar alentada por los Gobiernos, por ejemplo para orientar el comportamiento hacia una dirección más amistosa respecto del medioambiente.

Este tipo de intervenciones no eliminan la libertad de elección o de acción de las personas. No hay coerción. Este es el motivo por el que Thaler y Sunsteins contemplan el *nudge* como una forma de «paternalismo libertario» (5). No se realiza en contra de la voluntad de las personas. Por lo tanto, difiere del paternalismo clásico como, por ejemplo, lo definió Dworkin (2020): «la interferencia de un Estado o un individuo en otra persona, en contra de su voluntad, y defendida o motivada por la afirmación

de que la persona sobre la que se practica la interferencia estará mejor o más protegida del daño». Mientras que el paternalismo clásico viola claramente la libertad negativa, el *nudge* dirige las elecciones de las personas hacia lo que se supone que son sus mejores intereses sin restringir su libertad de la forma en la que Dworkin lo describe. O, usando la terminología de Berlin: el *nudge* no viola la libertad negativa, ya que no hay una restricción externa. Por ejemplo, un Gobierno que quiera promover la salud pública de sus ciudadanos puede necesitar que las compañías tabacaleras pongan un aviso en los paquetes de cigarrillos sobre que fumar mata, o que los supermercados los quiten de lugares destacados como las líneas de caja. Esta política no prohíbe los cigarrillos, pero necesita que los productores y distribuidores pongan *nudges* en ciertos lugares para que influyan en las elecciones de las personas. De forma similar, un sistema que haga recomendaciones y esté dirigido por una IA, no te fuerza a comprar un libro específico o a escuchar una canción concreta, pero puede influir en tu comportamiento.

Sin embargo, a pesar de que esto no es una amenaza a la libertad negativa, ya que nadie está forzado a hacer o a decidir algo, el *nudge* realizado por una IA es una amenaza a la libertad positiva. Al operar sobre la psicología subconsciente de las personas, las manipula sin respetarlas como seres racionales que desean establecer sus propias metas y tomar sus propias elecciones. Las formas de manipulación subconsciente como la publicidad y la propaganda no son nuevas, pero el *nudge* pretende ser libertario e, impulsado por la IA, es probable que tenga una influencia generalizada. El *nudge* puede llevarse a cabo por empresas, pero también por el Estado, por ejemplo para lograr que se alcance una sociedad mejor. Sin embargo, junto con Berlin (1997), se podría argumentar que las violaciones de la libertad positiva en nombre de una reforma social son degradantes: «Manipular a los hombres, para impulsarlos hacia las metas que tú (el reformador social) ves, pero puede que ellos no, es negarles su esencia humana, tratarles como objetos sin voluntad propia y, por lo tanto, degradarles» (209). De acuerdo con Berlin, el paternalismo es «un insulto a mí mismo como ser humano» (228) porque me falta al respeto como un ser autónomo que quiere tomar sus propias elecciones y dar forma a su propia

vida. Esta acusación parece ser también aplicable al llamado paternalismo «libertario» del *nudge*; en ese caso la gente no es siquiera consciente de que sus elecciones están siendo influenciadas, por ejemplo en el supermercado.

Esto convierte al *nudge* por medio de la IA en algo, como mínimo, altamente sospechoso y (como sucede con todas las violaciones de la libertad negativa) *prima facie* injustificado (injustificado a menos que se pruebe lo contrario). Alguien que aún quisiera usar la tecnología de esta forma tendría que argumentar que hay un principio más importante y mejor que la libertad positiva. Por ejemplo, se podría argumentar que la salud y la vida de una persona son más importantes que respetar su autonomía, o que la supervivencia de la humanidad y de otras especies es más importante que la libertad positiva de las personas, que son ignorantes de los efectos que producen sobre el clima o reacias a contribuir a resolver este problema. Pero, mientras que las presuntas vulneraciones de la libertad positiva son consideradas como menos controvertidas que las violaciones de la libertad negativa, es importante comprender lo que aquí está en juego: el riesgo de tratar a la gente como objetos que pueden y deben ser manipulados por su propio bien, o por el bien de la sociedad (por ejemplo, haciendo *nudge* contra la obesidad), sin tener en cuenta sus capacidades para la elección autónoma y la toma racional de decisiones, y tratarlos como medios para lograr determinados fines (por ejemplo, medios para alcanzar las metas climáticas), fines que otros (por ejemplo, los Gobiernos, los reformadores ecologistas) conciben de forma independiente de esa misma gente. ¿Cuándo y por qué los fines (por ejemplo, los fines deseables) justifican los medios, justifican tales degradaciones, si es que lo hacen? Y ¿quién decide los objetivos?

Concebir a la gente como principal o básicamente irracional y sorda a la argumentación es también una visión muy pesimista de la naturaleza y de la sociedad humanas, en la línea de la filosofía política de Hobbes. De acuerdo con Hobbes (1996), que escribió en la Inglaterra de mediados del siglo XVII, el estado de naturaleza sería una condición desagradable en la que solo habría competición y violencia. Para evitar esta situación, argumentaba, se necesitaba una autoridad política, un leviatán, para poner orden. De forma similar, el paternalismo libertario es pesimista en lo que concierne a las

capacidades de la gente para construir un orden social que sea bueno para ella misma y para la sociedad. Este orden social tiene que ser impuesto desde arriba mediante la manipulación, usando, por ejemplo, la IA. Otros filósofos políticos, como Rousseau en el siglo XVIII y Dewey y Habermas en el siglo XX, sostienen, por el contrario, una visión más optimista de la naturaleza humana y creen en formas democráticas de política de acuerdo con las cuales la gente es considerada capaz de comprometerse voluntariamente en pos del bien común, de deliberar racionalmente y de debatir en busca del consenso. Según estas opiniones, los humanos no deberían estar controlados (restricciones a la libertad negativa, despotismo) ni manipulados (restricciones a la libertad positiva, paternalismo). Son perfectamente capaces de contenerse, de pensar racionalmente y más allá de su propio interés, y de discutir entre ellos sobre lo que es bueno para la sociedad. En esta visión, la sociedad no es una colección de individuos atomizados, sino una república de ciudadanos dirigida hacia la consecución del bien común. Al igual que otros que defienden el *republicanismo* filosófico, Rousseau (1997) dirigía su atención hacia las antiguas ciudades-Estado griegas y pensaba que el bien común se podría alcanzar a través de una ciudadanía y participación activas, y que los ciudadanos deberían someterse a una «voluntad general», formando así una comunidad de iguales. Mientras que en el caso de Rousseau existe un conocido problema con su opinión de que uno debería ser *forzado a ser libre*, esto es, forzado «a consultar a la [propia] razón antes de escuchar las [propias] inclinaciones» (53) y a someterse a la voluntad general, rechazaba el despotismo y, en general, su visión es optimista en lo que respecta a la naturaleza humana: el estado de naturaleza es bueno y ya es social. También estaba de acuerdo con los filósofos de la Antigüedad, como Platón y Aristóteles, en que, a nivel personal, la libertad como autonomía es tanto alcanzable como deseable a través del uso de la razón y el control de las pasiones (una visión que también podemos encontrar en la ética de la virtud). Queda todavía abierta la cuestión de cuál es el papel que podría tener la IA bajo estos ideales. En el capítulo 4 esbozaré diferentes opiniones sobre la posibilidad de la democracia y de los ideales relacionados, y en el

capítulo 5, que trata sobre el poder, diré más sobre la IA y la autoconstitución.

AMENAZAS A LA AUTORREALIZACIÓN Y A LA EMANCIPACIÓN: LA EXPLOTACIÓN MEDIANTE EL USO DE LA IA Y EL PROBLEMA DE LOS ROBOTS ESCLAVOS

Otra amenaza a la libertad proviene no de la interferencia con la libertad negativa de cada uno o del uso de *nudges*, sino de violaciones de un tipo de libertad diferente, más relacional: la opresión y la explotación llevada a cabo por otros a través del trabajo en un contexto capitalista, o incluso (abiertamente) mediante la obligación impuesta a otras personas a soportar una relación de esclavitud y dominación. Mientras que esto puede involucrar restricciones a la libertad negativa de los individuos (por supuesto, no puedo hacer lo que quiero cuando soy un esclavo y *ni siquiera alcanzo una posición en la que importe lo que yo quiera*, ya que, para empezar, no estoy considerado como un sujeto político) y, mientras que la opresión también puede estar combinada con violaciones de la libertad positiva (explotación), estos fenómenos también plantean problemas relacionados con la *autorrealización*, el *autodesarrollo* y la *emancipación*, que enlazan con problemas de justicia e igualdad (véase también el siguiente capítulo), conciernen a la calidad de las *relaciones sociales* humanas, el valor del trabajo y su relación con la naturaleza y la libertad, así como la cuestión de cómo estructurar la sociedad. El tipo de libertad aquí amenazada es relacional en el sentido de que no trata sobre la gestión de deseos internos o sobre la consideración de los demás como amenazas externas, sino sobre la construcción de relaciones sociales y de sociedades mejores.

Hegel y Marx son fuentes de inspiración para esta concepción de la libertad. De acuerdo con Hegel, la transformación de la naturaleza a través del trabajo conduce a la autoconsciencia y a la libertad. Esta cuestión se remonta a la famosa dialéctica del amo y el esclavo de la *Fenomenología del espíritu* (1977; original de 1807): mientras el amo es dependiente de sus deseos, el esclavo alcanza la consciencia de la libertad a través del trabajo. Marx tomó prestada esta idea de que el trabajo conduce a la libertad. En sus

manos, la libertad ya no es una noción individual que tiene que ver con estar libre de restricciones o tener autonomía psicológica, como sucede con las libertades negativa y positiva de Berlin; es un concepto más social, materialista e histórico. Para Hegel y Marx, la libertad tiene sus cimientos en la interacción social. No es lo opuesto a la dependencia. El trabajo y las herramientas extienden nuestra libertad. Esta libertad tiene una historia, que es también una historia social y política y (se podría añadir) una historia de la tecnología. Marx pensaba que mediante la tecnología podemos transformar la naturaleza y, al mismo tiempo, crearnos a nosotros mismos. Mediante el trabajo, nos desarrollamos y ejercitamos nuestras capacidades humanas.

Sin embargo, Marx también defendía que, bajo el capitalismo, esto se vuelve imposible de alcanzar, ya que los trabajadores están alienados y explotados. En lugar de emanciparse y de realizarse, los trabajadores dejan de ser libres porque pasan a estar alienados de su producto, de los demás y, en última instancia, de sí mismos. En sus *Manuscritos económicos y filosóficos de 1844*, Marx (1977, 68-69) escribió que los propios trabajadores se convierten en productos y sirvientes de los objetos que producen y de aquellos que se apropian de esos productos. En lugar de afirmarse a sí mismos, los trabajadores mortifican sus cuerpos y arruinan su mente, convirtiendo su trabajo en algo forzoso en lugar de en algo libre (71). Se enajenan de sí mismos a medida que el trabajo se vuelve «una actividad desarrollada al servicio, bajo el dominio, la coacción y el yugo de otro hombre» (76). Bajo tales condiciones, la tecnología no conduce a la libertad, sino que se convierte en un instrumento de alienación. El comunismo, por el contrario, argumenta Marx, será la realización de la libertad, entendida como (de nuevo) la autorrealización y la asociación de personas libres.

¿Qué significa esta noción de libertad para la IA y la robótica?

Primero, la IA y sus dueños necesitan datos. Como usuarios de redes sociales y otras aplicaciones que necesitan nuestros datos, somos los trabajadores que producen esos datos. Fuchs (2014) ha defendido que las redes sociales y los motores de búsqueda, como Google, no son liberadores del capitalismo, sino que están colonizados por él. Estamos realizando

trabajo gratis para las empresas de redes sociales y sus clientes (anunciantes): producimos un bien (datos) que es vendido a ciertas empresas. Esta es una forma de explotación. El capitalismo necesita que trabajemos y consumamos continuamente, incluyendo aquí nuestro uso de dispositivos electrónicos para producir datos. La mayoría de nosotros vive en una economía capitalista 24/7: la única «libertad» que podemos encontrar es en el sueño (Crary, 2014; Rhee, 2018, 49). Incluso cuando estamos en la cama el teléfono exige nuestra atención. Además, los aparatos que utilizamos están producidos a menudo bajo «condiciones similares a la esclavitud» (Fuchs, 2014, 120), ya que dependen del duro trabajo de aquellos que los producen y que extraen los minerales con los que se fabrican. Los servicios de la IA también dependen de trabajadores mal pagados que limpian y etiquetan los datos, ensayan modelos, y así sucesivamente (Stark, Greene y Hoffmann, 2021, 271). Sin embargo, a la luz del análisis de Marx en términos de libertad como autorrealización, el problema que surge cuando usamos las redes sociales no es solamente que estemos realizando trabajo gratuito y que otros estén siendo explotados para permitir nuestros placeres en las redes sociales (que pueden ser analizados sobre la base del enfoque económico político de *El capital* de Marx [1990, originalmente, 1867]: el valor creado por los trabajadores que excede el coste de su propio trabajo se lo apropia el capitalista), sino también que no está conduciendo a nuestro autodesarrollo y autorrealización y, por ende, a nuestra libertad. Al contrario, nos convertimos en objetos: un conjunto de datos (ver también el capítulo 5).

Segundo, los robots son una tecnología que es empleada a menudo para la automatización, y esto conlleva efectos que se pueden describir usando el concepto marxista de la libertad humana. Para empezar, los robots tienen la forma de máquinas que contribuyen a la alienación descrita por Marx: los trabajadores se convierten en una mera parte de la máquina y pierden la oportunidad de realizarse a través del trabajo. Esto ya ha ocurrido en la producción industrial: pronto ocurrirá también en la industria de servicios, por ejemplo en el comercio minorista o en restaurantes (por ejemplo, en Japón). Además, como describe Marx, el uso de las máquinas conduce no solo a malas condiciones laborales y al deterioro físico y mental de los

trabajadores, sino también al desempleo. El uso de robots para reemplazar a los trabajadores humanos crea una clase de personas desempleadas que solo tienen para vender su propio trabajo (el proletariado). Esto no es malo solamente para aquellos que pierden el empleo, sino que también hace decrecer los salarios de aquellos que aún continúan empleados (o se mantienen los salarios tan bajos como lo permite la legalidad). Además, algunos ven que su trabajo se devalúa: también podría hacerlo un robot (Atanasoski y Vora, 2019, 25). El resultado es la falta de libertad en el sentido de explotación y la carencia de oportunidades para la autorrealización.

Mientras que actualmente se acepta ampliamente que la robótica y la IA tendrán muy probablemente consecuencias en el empleo (Ford, 2015), los autores no están de acuerdo sobre la velocidad prevista y el alcance de estos desarrollos. Economistas como Stiglitz predicen una importante alteración y advierten sobre el coste humano de la transición. Por ejemplo, Korinek y Stiglitz (2019) predicen una importante perturbación de los mercados laborales, que llevaría a una mayor desigualdad salarial, mayor desempleo y una sociedad más dividida, a menos que los individuos se encuentren completamente asegurados contra estos efectos y se hayan implementado formas correctas de redistribución (tales como las características de las democracias del bienestar social europeas). Además, las consecuencias socioeconómicas de la IA pueden variar entre las denominadas sociedades avanzadas y el hemisferio sur (Stark, Greene y Hoffmann, 2021). Desde un punto de vista marxista, estas cuestiones pueden ser conceptualizadas en términos de igualdad, pero también desde la libertad como autorrealización. Los bajos sueldos y el desempleo no son solamente malos porque amenazan la existencia física de las personas: también hacen que la gente sea políticamente menos libre, ya que no se pueden realizar a sí mismos.

En contra de tales opiniones, algunos han aducido que las máquinas liberarán a la humanidad del trabajo sucio, pesado, peligroso o aburrido, dejando tiempo libre para el ocio y la autorrealización. El desempleo debido a la sustitución de las personas por las máquinas es bienvenido como un paso en el camino hacia la libertad. En lugar de considerar el trabajo como un camino hacia la libertad, como hacían Hegel y Marx, esta visión adopta

la antigua idea aristotélica de que la libertad consiste en liberarse a uno mismo de las necesidades vitales. De acuerdo con Aristóteles, ocuparse de las necesidades vitales es algo propio de esclavos, no de gentes libres. Los marxistas, sin embargo, no solamente estarían en desacuerdo con esta opinión sobre el trabajo, sino que también señalarían que la sociedad de Aristóteles se basaba en la esclavitud: la elite política solo podría disfrutar de sus vidas privilegiadas explotando a los demás. Un defensor de la que normalmente es denominada «la sociedad del ocio» podría responder que la tecnología acabaría con la esclavitud humana y que las consecuencias del desempleo podrían tratarse mediante un sistema de seguridad social, mediante, por ejemplo, una renta básica universal que garantizaría que nadie, incluyendo los desempleados por mor de las máquinas, fuese pobre. Pero los marxistas podrían aducir que el capitalismo de la IA podría dejar de necesitar a los humanos en todos los aspectos: en este escenario, el capital gana «la libertad de una humanidad que se convierte en una barrera biológica contra la acumulación» (Dyer-Witheford, Kjosen y Steinhoff, 2019, 149).

Estas cuestiones hacen que surjan nuevas preguntas filosóficas. Desde una perspectiva hegeliana, por ejemplo, nos podría preocupar el que si todos los humanos se convirtieran en amos (amos de las máquinas), estos carecerían de oportunidades para la autorrealización y se entregarían a sus deseos, que estarían manipulados y explotados por los capitalistas. Los amos, a su vez, se convertirían en consumidores explotados: un nuevo y diferente tipo de esclavos. Hasta cierto punto, este parece ya ser el caso actualmente. Como ha argumentado Marcuse (2002), la sociedad de consumo produce nuevas formas de dominación. No solo los amos son dependientes de las máquinas que controlan: como consumidores, están, de nuevo, dominados. Además, continuando con la dialéctica del amo y el esclavo, una vez que se sustituyan los esclavos por máquinas, los amos (en este escenario: todos nosotros) ya no tendrán la oportunidad de reconocerse como tales, ya que no podremos recibir reconocimiento alguno por parte de las máquinas, puesto que estas carecen de la necesaria autoconciencia. Así, si Hegel está en lo correcto al afirmar que el amo depende del reconocimiento por parte del esclavo, entonces el problema es que, en esta

situación, el amo no recibe ningún tipo de reconocimiento. En otras palabras, en una sociedad de amos humanos y robots esclavos, los amos-consumidores no son en absoluto libres y, peor aún, ni siquiera tienen la oportunidad de serlo. Como esclavos-consumidores, se encuentran dominados y explotados bajo el capitalismo. Como amos de máquinas, no reciben reconocimiento. Y, como he subrayado en trabajos anteriores sobre este tema, también se vuelven altamente dependientes de la tecnología y, por ende, vulnerables (Coeckelbergh, 2015a).

Pero ¿por qué pensamos en términos de amos y esclavos? Por un lado, reemplazar a los esclavos o trabajadores humanos por robots puede considerarse como liberador, desde un punto de vista marxista e, incluso, desde uno más general, ilustrado: los humanos ya no están insertos en estas relaciones sociales de explotación. Por otro lado, pensar en términos de siervos y esclavos, incluso en el caso de robots, parece bastante problemático. ¿Está bien tener robots esclavos? ¿Está bien *pensar* sobre una sociedad en términos de amos y esclavos? Y no se trata simplemente de pensar. Si reemplazamos a los esclavos humanos por máquinas la estructura de nuestra sociedad seguirá estando basada en la esclavitud, a pesar de que los esclavos sean artificiales: una versión cibernética de la antigua Roma o de las antiguas ciudades-Estado griegas. No hay ninguna violación del artículo 4 de la DUDH (Declaración Universal de los Derechos Humanos), ya que el artículo se aplica solo a los humanos (trata sobre derechos *humanos*) y los robots no son esclavos en el sentido humano, ya que carecen de consciencia, sensibilidad, intencionalidad, etc. Pero la configuración en la que los robots ocupan el puesto de los sirvientes o esclavos humanos refleja aún el pensamiento amo-esclavo y su sociedad jerárquica y «explotadora». Lo que está mal es el tipo de relaciones sociales y el tipo de sociedades que ayudan a mantener estas tecnologías (el siguiente capítulo ofrece un argumento adicional).

Ahora bien, se podría intentar sortear estos problemas argumentando que no se trata de libertad, sino de algo diferente. El problema de la opresión, la explotación y la esclavitud no tiene que ver con la *libertad* (al menos, si se entiende la libertad de la forma individualista y más formal articulada en las secciones anteriores), sino con la *justicia* o la *igualdad*. De acuerdo con esta

concepción, el problema no es que la IA y la robótica amenacen la libertad, sino que vivimos en una sociedad que es fundamentalmente desigual o injusta y que nos arriesgamos a mantener o empeorar estas desigualdades e injusticias mediante la IA y la robótica. Por ejemplo, si «liberamos» a los trabajadores reemplazándolos por máquinas, pero no cambiamos la estructura de nuestras sociedades actuales (mediante una renta básica universal u otras medidas), es muy probable que creemos más desigualdad e injusticia. Necesitamos más discusión pública sobre qué tipo de sociedades tenemos y queremos. Los conceptos y teorías político-filosóficos tales como los de igualdad y justicia pueden ayudarnos en este trance. Por ejemplo, como en los sistemas de seguridad social actuales, la renta universal básica refleja una concepción concreta de la justicia distributiva y de la justicia como equidad. Pero ¿cuál de ellas? En el siguiente capítulo, esbozaré algunas concepciones de la justicia.

Sin embargo, es también interesante discutir estas cuestiones desde la óptica de la libertad. Por ejemplo, en *Libertad real para todos* (1995), Van Parijs defiende una renta básica incondicional para todos basada en las nociones de justicia, igualdad y *libertad*. Entiende la libertad no como un derecho formal para hacer lo que uno quiera (la noción libertaria usual de la libertad tal como aparece, por ejemplo, en Hayek), sino como la capacidad real de hacerlo. Así, la libertad se define en términos de oportunidad. Las personas con más ventajas (por ejemplo, mayor acceso a los recursos) tienen más libertad que otras en términos de oportunidad. Una renta incondicional haría que los menos favorecidos fuesen más libres en este sentido de oportunidad, a la vez que respeta la libertad formal de los demás, creando de este modo libertad para todos. Además, todos podrían usar estas oportunidades de acuerdo con su propia concepción de lo que es bueno. Los liberales deberían ser neutrales con respecto a las concepciones personales de lo que es bueno (dentro de los límites de la legalidad). Tomando uno de los ejemplos de Van Parijs: si las personas quieren pasar una gran parte de su tiempo surfeando, no hay problema: tienen la oportunidad de hacerlo. Van Parijs llama a esto «libertarismo real». Sobre la base de esta perspectiva, se podría decir que, cuando las máquinas se hagan cargo de los trabajos de los humanos, la renta universal básica sería no solo una forma

de crear más justicia e igualdad, sino también de respetar y promover la libertad de todos. La gente podría trabajar, o surfear, o ambas cosas: tendrían libertad real en el sentido de libertad de oportunidad.

Sin embargo, como muestra la discusión sobre la renta universal básica, hay mucho más que decir sobre las dimensiones políticas y sociales de la IA y de la robótica en términos que van más allá de la libertad. También necesitamos hablar de la igualdad y de la justicia, y sobre cómo la IA y la robótica pueden perpetuar o agravar las formas existentes de sesgo y discriminación (ver el capítulo siguiente).

¿QUIÉN DECIDE SOBRE LA IA? LIBERTAD COMO PARTICIPACIÓN, IA EN LAS ELECCIONES Y LIBERTAD DE EXPRESIÓN

Otro significado de la libertad es la *participación política*. De nuevo, esta idea es antigua. Tiene, de hecho, raíces aristotélicas. Como explica Arendt en *La condición humana* (1958), la libertad para los antiguos no era la libertad liberal de elección, sino la de acción política, que Arendt distingue de otras actividades en la *vita activa* (ocupaciones personales y trabajo). De acuerdo con el republicanismo filosófico, solo se puede ejercer la propia libertad mediante la participación política. Mientras que para los antiguos griegos esta libertad estaba reservada para una elite y, de hecho, dependía de la esclavitud de aquellos que trabajaban y a los que, por tanto, se les negaba la libertad política, la idea de libertad como participación política es importante en la historia de la filosofía política moderna y ha inspirado muchas interpretaciones e ideales clave de la democracia (ver capítulo 4). Como ya indiqué anteriormente en este capítulo, una expresión moderna bien conocida de la idea de libertad como participación en la toma de decisiones políticas proviene de Rousseau, quien argumentaba ya antes que Kant que la libertad significa autorregularse. Esta autorregulación puede ser interpretada como autonomía individual (ver la sección más arriba sobre el *nudge*), pero Rousseau también otorgó a la autorregulación un significado político: si los ciudadanos crean sus propias leyes, son realmente libres y no se encuentran a merced de la tiranía de otros. A pesar de lo controvertido que puedan ser los pensamientos adicionales de

Rousseau sobre la voluntad general, la idea de que la participación política es y debe ser parte de la democracia liberal ha cuajado e influido en la forma en que muchos de nosotros pensamos sobre la democracia hoy en día.

¿Qué significado tiene la concepción de la libertad como autoparticipación para la cuestión de la libertad en relación con la IA y la robótica? He aquí algunos argumentos normativos que se podrían poner sobre la mesa respecto de la política de la IA y la robótica que están basados en esta concepción.

Primero, a menudo las decisiones sobre las tecnologías y su uso no las toman los ciudadanos, sino los Gobiernos y las corporaciones: los políticos, gestores, inversores y los que desarrollan la tecnología. Este también es el caso de la IA y la robótica, que se desarrollan frecuentemente en contextos militares (financiados por el Gobierno) y en empresas tecnológicas. Basándonos en el ideal de la libertad como participación política, se podría criticar este hecho y exigir la participación de los ciudadanos en las discusiones públicas y políticas sobre la IA y la robótica. Mientras que se puede argumentar de esta manera basándose en el principio de la democracia, desde la perspectiva del republicanismo filosófico tal argumentación puede estar justificada sobre la base de la libertad: si la libertad implica participación y autogobierno políticos, entonces la situación actual en la que los ciudadanos tienen poca, o ninguna, influencia sobre el uso y el futuro de la tecnología les está manteniendo de hecho en un estado de falta de libertad y despotismo. Los ciudadanos carecen de autogobierno respecto de esas decisiones. En nombre de la libertad como participación se podría reclamar una toma de decisiones más democrática en lo que respecta a nuestro futuro tecnológico.

Además, más allá de exigir cambios a las usuales instituciones políticas, se podría pedir que los propios procesos de innovación tomen en cuenta a los ciudadanos y a sus valores como parte interesada que son. Durante la década pasada hubo discusiones sobre la innovación responsable y el diseño sensible a los valores (Stilgoe, Owen y Macnaghten, 2013; van den Hoven, 2013; von Schomberg, 2011): la idea es que los actores sociales están incluidos en procesos de innovación y que los valores éticos se tienen en

cuenta durante la fase de diseño. Pero esta no es solamente una cuestión de responsabilidad ética: basándose en el ideal de libertad como participación, se podría ver también como un imperativo *político*. La libertad no debería ser solamente mi libertad de elección como usuario o consumidor de tecnología (considerando la tecnología que uso como dada), sino también sobre mi libertad para participar en decisiones y en procesos de innovación que conciernen a las tecnologías que uso. Ejercer esta libertad es especialmente importante, ya que, como continúan insistiendo los filósofos de la tecnología, las tecnologías tienen efectos no intencionados y conforman nuestras vidas, sociedades y las personas que somos. Si la libertad como autogobierno es un principio político importante, entonces, con respecto a tecnologías como la IA, es importante no solo que adquiera autonomía personal sobre mi uso de la tecnología (y responsabilidad) como usuario y consumidor, sino también que yo, como ciudadano, tenga algo que decir sobre las decisiones relacionadas con la tecnología y responsabilidad política sobre ellas. Si falta este tipo de libertad como participación, entonces el futuro de la IA y de la robótica (que también es *mi* futuro) permanece en las manos de los políticos tecnocráticos y de los CEO, dueños e inversores despóticos.

Segundo, sin embargo, incluso como participantes en la política podemos estar siendo manipulados mediante la IA. La IA desempeña un papel cada vez más importante en las campañas electorales y en la vida política en general. Por ejemplo, la evidencia sugiere que la IA se usó para manipular a los ciudadanos en la campaña presidencial de Donald Trump de 2016 (Detrow, 2018): la empresa de ciencia de datos Cambridge Analytica envió publicidad personalizada a votantes individuales basándose en la elaboración del perfil de su psicología, usando datos sobre sus comportamientos en redes sociales, patrones de consumo y relaciones sociales. Se pueden utilizar bots en redes sociales como Facebook y Twitter, disfrazados como cuentas humanas, para propagar desinformación y noticias falsas entre grupos demográficos específicos (Polinski, 2017). Este es un problema para la democracia (ver capítulo 4), pero también para la libertad. Se relaciona de nuevo con amenazas a la libertad como autonomía

y a la cuestión de la vigilancia, pero también influye en la libertad como participación política.

No obstante, el problema de la manipulación no es solamente sobre la manipulación política en sentido estricto: esto es, la manipulación para propósitos políticos en contextos que normalmente son definidos como políticos. Cuando usamos por doquier nuestros dispositivos inteligentes en el trabajo, pero también en casa, estamos operando en entornos cada vez más inteligentes en los que los agentes automatizados moldean nuestros entornos de elección. Hildebrandt (2015) considera esta cuestión en analogía con el sistema nervioso autónomo que gobierna nuestro ritmo cardíaco: de forma similar al sistema nervioso autónomo que ajusta nuestro entorno interno (nuestro cuerpo), un sistema de ordenador autónomo puede ahora ajustar nuestros entornos externos «para hacer lo que infiere que es necesario o deseable para nuestro propio bienestar». Y todo esto ocurre sin nuestro conocimiento consciente, de la misma manera que no sabemos cómo están gobernadas nuestras funciones vitales (55). Este es un problema para la libertad como autonomía y para la sociedad en su conjunto, siempre y cuando «el entorno social deje de estar formado por medio de mutuas expectativas entre las personas y pase a estar dirigido exclusivamente por las metas de la manipulación dirigida por datos» (Couldry y Mejias, 2019, 182). Pero esto también nos hace pensar sobre qué tipo de participación política queremos y necesitamos. ¿Cómo podemos volver a adquirir más control sobre lo que nos ocurre a nosotros y a nuestros entornos, tecnológicamente hablando? ¿Cómo podemos alcanzar el autogobierno propuesto por el republicanismo filosófico y el pensamiento ilustrado?

En el capítulo 4 diré más sobre la democracia como participación, pero en relación con la libertad se puede argumentar que una condición para la libertad como participación política es la educación. Rousseau no defendería un sistema político como el nuestro, en el que la participación política se encuentra dissociada de la educación. En su lugar, y como hizo Platón, Rousseau propuso una educación moral para los ciudadanos. Para él, esta es la única manera en la que podemos lograr el ideal de libertad como participación política. Se rebelaría contra la idea de que lo único que necesitan hacer los ciudadanos es votar cada cuatro o cinco años y que el

resto del tiempo se les deje hacer lo que quieran, encontrándose, como efectivamente están, en manos de las redes sociales. Él y otros pensadores ilustrados también quedarían horrorizados ante la idea de que los ciudadanos son un tipo de consumidores en el marco de una «política de autoservicio» y e-gobierno en el que la Administración pública está «orientada hacia el consumidor» (Eriksson, 2012, 691), o incluso que son coproductores de «servicios públicos» (687), a pesar de que lo último es ciertamente participativo y otorga a los ciudadanos un papel más activo (691). En su lugar, Rousseau, como Platón, pensaba que la educación haría que la gente fuera menos egoísta, más compasiva y menos dependiente de los demás, conduciéndonos hacia la dignidad moral y el respeto y, como dice Dent (2005, 150), dando lugar al «establecimiento de su completa humanidad en sus relaciones recíprocas». Esta educación moral y política llevaría a los ciudadanos a obedecer la voluntad general, en el sentido de «hacer justamente lo que uno haría como ser moral» (151). Este ideal de libertad política como libertad moral basada en la educación se encuentra amenazado por una esfera pública dominada por la manipulación y la desinformación, y la IA puede desempeñar un papel importante a la hora de crear y mantener tal esfera de poder.

Esto nos conduce a otra pregunta importante para la libertad en esta área: ¿deberían las redes sociales estar más fuertemente reguladas, o deberían autorregularse para crear una discusión pública y una participación política de mayor calidad a expensas de perder alguna porción de la libertad negativa? O ¿es la libertad negativa, en la forma de libertad de opinión y *libertad de expresión*, entendida como libertad para decir lo que se quiera, más importante que la libertad como participación y acción políticas? ¿Es la difusión de la desinformación y el odio una forma genuina de participación y acción políticas? ¿Es aceptable, ya que respeta la libertad negativa (libertad de expresión), o es lo opuesto a la libertad, ya que puede conducir al totalitarismo (y, por tanto, a la carencia de libertad negativa) y a la propia corrupción de la política y de la libertad como participación? Basándose en Aristóteles y Rousseau, se podría apoyar la última opinión, criticar los puntos de vista liberales (o, al menos, sus interpretaciones libertarias) y argumentar que la libertad de expresión no incluye la libertad de decir cosas

que busquen destruir la democracia, y que los ciudadanos deberían ser educados para convertirse en mejores seres morales y participar en la política de una forma que realice su humanidad. Mientras que el primer argumento está bastante consolidado en nuestras democracias actuales como justificación del establecimiento de límites a la libertad de expresión, el segundo (la educación moral y la participación política de los ciudadanos) es mucho más controvertido. Se necesitaría una regulación de la esfera pública mediada por la IA y otras tecnologías digitales, junto con importantes reformas en educación y en las instituciones políticas.

Las restricciones a la libertad de expresión pueden estar impuestas por humanos, pero también por la IA. Las plataformas de redes sociales digitales como Twitter e incluso medios más tradicionales (por ejemplo, los foros de discusión *online* de los periódicos) ya usan la IA para la llamada «moderación de contenido», que puede implicar la detección automática de contenido potencialmente problemático y la eliminación automática o la dulcificación de ese contenido. Estas acciones se pueden utilizar para bloquear opiniones que se consideren problemáticas o para eliminar desinformación y noticias falsas (por ejemplo, en forma de textos o vídeos) que se utilizan para manipular políticamente a las audiencias. En respuesta a este uso de la IA, nos podríamos preguntar cuán precisa es esta evaluación (en comparación con la evaluación humana) y si la evaluación humana es correcta. Existe también la preocupación de que, cuando falta un juicio humano, se pueda dar rienda suelta al uso de la IA para propagar la desinformación, y que automatizar las decisiones editoriales haga surgir problemas de responsabilidad (Helberg *et al.*, 2019).

Pero ¿qué es lo que falta exactamente? Esta cuestión se relaciona con un debate más amplio sobre el juicio humano contra el «juicio» de la IA, al que nos podemos aproximar ahora usando ciertas discusiones sobre el juicio *político*. Por ejemplo, de acuerdo con Arendt (1968, interpretando la teoría estética de Kant), el juicio político tiene que ver con un *sensus communis*, o sentido común, con tener un mundo en común (compartir un mundo) y con el uso de la imaginación para tomar en cuenta los puntos de vista de los demás. Arendt también se refiere a la noción de *frónesis* de Aristóteles, entendida a menudo como sabiduría práctica. Esta noción es conocida en la

ética de la virtud, que se centra en el carácter y los hábitos morales de las personas y que se ha utilizado para reflexionar sobre la robótica (por ejemplo, Coeckelbergh, 2021; Sparrow, 2021), pero en la obra de Arendt también desempeña un papel político. De acuerdo con ella, el juicio político exige deliberación e imaginación. Quizás incluya también un componente afectivo, como ha argumentado Aavitsland (2019), influenciado por Arendt: una afirmación que saca partido de la antigua discusión sobre el rol de la racionalidad y de las emociones en la política. ¿Cómo podría la IA llegar a tener tal capacidad para el juicio político, dado que, careciendo de consciencia y no siendo parte de ningún «mundo» en ese sentido, no tiene ninguna subjetividad, no digamos ya intersubjetividad, imaginación, emociones, etc.? Por otro lado, ¿cómo son de buenos los humanos al hacer juicios políticos? Claramente, no consiguen a menudo ejercitar este tipo de sentido común, imaginación política y juicio y no van más allá de defender sus intereses privados. Además, aquellos que (contra Arendt) mantienen una noción racionalista y naturalista del juicio, pueden estar tentados a argumentar que la IA podría hacerlo *mejor* que los humanos y ofrecer un juicio más «objetivo», libre de emociones y sesgos. Algunos transhumanistas discuten escenarios en los que tal IA asume el poder que tienen hoy día los humanos y sugieren que este sería un buen próximo paso en la historia de la inteligencia. Pero ¿es posible que haya en realidad un juicio objetivo o libre de sesgos? Después de todo, la IA puede también estar sesgada algunas veces, si no lo está siempre. Volveré sobre esta cuestión en el capítulo siguiente. Y, por supuesto, como los humanos, la IA puede cometer errores: ¿estamos dispuestos a aceptar esto en la política, pero también, por ejemplo, en un contexto médico (por ejemplo, diagnosis, decisiones sobre vacunación) o en la carretera (coches autónomos)? En cualquier caso, una cosa es decir que en algunos contextos la IA puede ayudar a los *humanos* a juzgar, mostrando patrones en los datos y ofreciendo recomendaciones y predicciones basadas en el cálculo de probabilidades, y otra es afirmar que lo que hace la IA consiste en realizar un juicio.

Con todo, a la luz del tema de este capítulo, la cuestión principal que discutir aquí tiene que ver con la libertad: ¿la moderación de contenido que

lleva a cabo la IA constituye un daño a la libertad de expresión y está realmente justificada? Hay muchas cautelas sobre esta cuestión. Llansó (2020) ha argumentado que, «a pesar de los avances en el aprendizaje automático, las órdenes de filtrado son una amenaza para la libertad de expresión» y, por lo tanto, para los derechos humanos. Por ejemplo, diversos defensores de los derechos humanos han llamado la atención sobre el impacto de la IA, e incluso en las Naciones Unidas un relator especial sobre la libertad de expresión, ha aplicado la ley de los derechos humanos para evaluar el impacto de la IA usada para moderar y organizar el contenido en plataformas digitales (UN, 2018). El informe establece una conexión con el artículo 19 de la DUDH. Además de recurrir al principio de la libertad de palabra en el sentido estricto de libertad de expresión, se podría argumentar que la gente tiene derecho a ser informada y a participar libremente en discusiones públicas. Por ejemplo, la UNESCO afirma promover el «libre flujo de ideas por medio de palabras e imágenes» (MacKinnon *et al.*, 2014, 7).

La libre discusión e intercambio de ideas es especialmente importante en una democracia liberal. Utilizando la filosofía política, se podría argumentar con Mill que eliminar la libertad de expresión pone en riesgo de asfixia el debate intelectual. En *Sobre la libertad* (1978, originalmente 1859), Mill defendió la libre expresión de opiniones sobre la base de que, de lo contrario, acabaríamos en una «pacificación intelectual» (31) en lugar de tener mentes que se atrevieran a llevar los argumentos a sus límites lógicos. Sin embargo, para Mill, la libertad de opinión no es absoluta, en el sentido de que debería impedírsele a la gente que haga daño a otras personas (el famoso principio del daño). En la tradición liberal anglófona, esto se interpreta normalmente como daño a los derechos individuales. Para Mill, se trata de no dañar a otros individuos y, en última instancia, de maximizar su felicidad. Sin embargo, en la tradición del republicanismo filosófico la cuestión no es que pueda hacerse daño a los individuos. El problema es que el discurso del odio, la manipulación y la desinformación ponen en peligro la libertad de participación y la autorrealización políticas, la dignidad moral de las personas y el bien común. El daño que se hace es, entonces, el daño a la posibilidad de la propia política. Y este

republicanismo filosófico parece compatible con al menos un aspecto del argumento de Mill: la importancia de disponer de libertad de expresión no reside en la expresión como tal, sino en la *argumentación* política y el debate *intelectual*. Esta es la libertad que debería respetarse. Si permitimos la libre expresión, deberíamos hacerlo en beneficio del respeto a la dignidad de las personas y (podemos añadir desde una perspectiva republicana) en beneficio de promover la discusión que debería conducirnos al bien común y a una ulterior realización de la humanidad.

De acuerdo con Mill, la discusión debería conducirnos a la verdad. Argumenta que la libre expresión es necesaria en aras de encontrar la verdad a través del debate. La verdad es valiosa, y las personas podrían estar equivocadas. Todo el mundo es falible. El libre mercado de ideas incrementa la posibilidad de que pueda emerger la verdad y de que se desafíen las creencias dogmáticas (Warburton, 2009). Sin embargo, la manipulación, la desinformación y la propaganda antidemocrática no conducen a estas metas. Ni conducen a un buen debate político (el tipo de participación política que deseamos, de acuerdo con Mill), ni apoyan las metas de la ilustración liberal ni de la moral republicana y el progreso político. Si el uso de la IA conduce a una esfera pública en la cual alcanzar estas metas se convierte en una tarea muy complicada, entonces la IA debería, o bien no usarse en esta área, o bien regularse de tal forma que apoye, en lugar de que corrompa, estos ideales políticos.

Sin embargo, si se quiere tener una regulación de este tipo, sigue siendo difícil especificar quién debería regular a quién, y qué salvaguarda de la libertad se necesita activar. Existe una importante discusión sobre el papel de las redes sociales. Por un lado, puede considerarse antidemocrático que las grandes compañías tecnológicas, en forma de empresas como Facebook y Twitter, decidan a quién debería censurarse y cuándo. De forma más general, se ha cuestionado por qué esas compañías tienen tanto poder. Por ejemplo, nos podríamos preguntar por qué tales plataformas, infraestructuras y editores necesitan estar en manos privadas, dado su importante papel en la democracia actual. Puesto que dominan el panorama de los medios de comunicación, los medios públicos de radiodifusión (configurados originalmente para apoyar la democracia por medio de la

educación) y los medios de comunicación tradicionales en general desempeñan ya un papel político importante. Desde esta perspectiva, la regulación es lo menos que se puede esperar. Por otro lado, si el Gobierno asume el papel de censor, ¿está esto justificado? y ¿qué criterios precisos debería usar? Los Gobiernos democráticos imponen restricciones a la libre expresión de forma rutinaria, pero como filósofos debemos preguntarnos tanto por la legitimidad de este rol como por los criterios usados. Y ¿qué pasa si se hace cargo del gobierno un régimen antidemocrático? Si dejamos que se erosione nuestra libertad de expresión ahora, será aún más sencillo para los regímenes autoritarios destruir la democracia.

Obsérvese también que los medios de comunicación tradicionales como los periódicos y la televisión, al menos los más o menos independientes y de alta calidad, ya tuvieron que lidiar con el problema de encontrar un equilibrio entre la libertad de expresión y la moderación para poder facilitar una buena discusión. Sin embargo, en el caso de las empresas de redes sociales privadas como Twitter y los periódicos tradicionales establecidos, el *qué*, *por qué*, *cómo* y *quién* de esta moderación y censura no es completamente transparente: a menudo no sabemos qué voces no son escuchadas, cómo se ha justificado esa decisión (si es que se ha justificado), qué procedimientos se siguieron (si es que se siguieron) y quiénes fueron los moderadores y censores. Además, vemos que, en la actualidad, los medios de comunicación de carácter nacional hacen un uso cada vez mayor de moderación automática de contenidos: justamente como en Facebook y en Twitter. Esto plantea problemas similares en relación con la libertad de expresión, junto con problemas de equidad. Mientras que la mayoría de la gente parece pensar actualmente que no hay una libertad de expresión absoluta y acepta *de facto* alguna moderación, sigue habiendo serios problemas, tanto en términos de libertad como en términos de otros valores y principios.

Por ejemplo, la teoría crítica contemporánea ha criticado el nexo de la filosofía liberal clásica entre la libertad de expresión y la libertad negativa entendida como libertad de interferencia por no prestar atención a la desigualdad, el poder, el racismo y el capitalismo estructurales (ver también el siguiente capítulo). Titley (2020) ha analizado cómo la libre expresión se

ha movilizado en la política de derechas para propagar ideas racistas, y cómo las concepciones superficiales de la libre expresión pasan por alto las desigualdades estructurales entre diferentes hablantes. Junto con el fenómeno de los bulos, esto ayuda a explicar por qué en el contexto político actual de los Estados Unidos muchos liberales exigen que las empresas como Facebook y Twitter impongan más regulación y restricción de expresión en casos de, por ejemplo, discursos de odio, ideas racistas y desinformación, y consideran las apelaciones a la libertad de expresión como problemáticas *a priori*. En respuesta a esto se podría argumentar que el principio de libertad de expresión como tal no es problemático, sino que lo que estos problemas muestran es que necesitamos una concepción diferente y más rica de este principio, especialmente en la era de la comunicación digital: una concepción que sea más inclusiva, que promueva la verdad, que sea pluralista y crítica y que tome en consideración las lecciones del pensamiento político contemporáneo, a la vez que mantenga algunos aspectos de la concepción de Mill, como la creencia en el debate intelectual libre. Necesitamos, entonces, continuar con la discusión sobre qué implica la libertad de expresión en la era del periodismo automatizado, la censura de la IA y los bulos (incluyendo aquí los vídeos que difunden embustes y otros productos de la IA).

OTRAS NOCIONES POLÍTICAMENTE RELEVANTES DE LA LIBERTAD Y OTROS VALORES

Existen más nociones de libertad que tienen impacto sobre la IA. Por ejemplo, transhumanistas (ver capítulo 6) como Bostrom, Sorgner y Sandberg han propuesto añadir la «libertad morfológica» a los derechos liberales individuales tradicionales. Aquí la idea es que a través de tecnologías avanzadas (IA, pero también nanotecnología, biotecnología, etc.) seremos capaces de remodelar la forma humana más allá de las limitaciones biológicas actuales y controlar nuestra propia morfología (Roden, 2015). Esto se podría entender como libertad al nivel de la humanidad en su conjunto, pero también como libertad individual. Por ejemplo, Sandberg (2013) ha argumentado que el derecho a modificar el

propio cuerpo es esencial para cualquier sociedad democrática del futuro y que, por lo tanto, debería ser considerado como un derecho básico. Se podría hacer una defensa similar del derecho a modificar la propia mente, por ejemplo con la ayuda de la IA. Aquí no se trata tanto de la concepción de libertad tradicional que tiene que ser revisada y discutida de nuevo a la luz de los recientes desarrollos tecnológicos: más bien sucede que las nuevas tecnologías nos otorgan un nuevo tipo de libertad del que no disponíamos previamente. Obsérvese, sin embargo, que tal libertad estaría muy cercana a la noción liberal clásica de libertad como autonomía (libertad positiva) y, especialmente, de la no interferencia (libertad negativa): la idea es que yo, y no otros, decido lo que hago con mi vida, mi cuerpo y mi mente.

Para concluir, Marx y la teoría crítica contemporánea, pero también el republicanismo filosófico inspirado por la filosofía aristotélica, desafían la idea libertaria liberal clásica y contemporánea de que la libertad tiene que ver principalmente con la no interferencia y la separación de los demás y con nuestra autonomía personal, definida de una forma (meramente) psicológica. Al contrario, defienden concepciones de la libertad más relacionales y políticas, de acuerdo con las cuales solo podemos ser realmente libres y emanciparnos si nos realizamos como seres políticos y nos convertimos en parte de una comunidad política de iguales que alcanzan el autogobierno y que crean las condiciones para la igualdad, la inclusión y la participación. Este enfoque nos permite realizar la pregunta sobre cuál es el papel de la IA (y sus usuarios humanos) y cuál puede ser a la hora de impedir o facilitar la realización de tal concepción de la libertad, la política y sus condiciones. Además, cada una de estas direcciones de pensamiento cuestiona, a su manera, la dirección y la organización de la sociedad en su conjunto en lugar de centrarse en el daño a los individuos. Si bien, a pesar de su uso indebido, la libertad de interferencia y el ejercicio de la autonomía personal interna continúan siendo principios y valores importantes en las democracias liberales actuales, las filosofías mencionadas ofrecen un marco alternativo interesante con el que discutir qué tipo de libertades políticas facilitan o amenazan la IA y la robótica y, ciertamente, qué tipo de libertades políticas y qué sociedad queremos.

Sin embargo, la libertad no es lo único que importa en la política de la IA. Como ya he sugerido, la libertad se encuentra también relacionada con otros valores políticos importantes, así como con principios y conceptos tales como democracia, poder, justicia e igualdad. Por ejemplo, de acuerdo con Tocqueville (2000), que escribió en la década de los 30 del siglo XIX, hay negociaciones y tensiones fundamentales entre la libertad y la igualdad. Advirtió sobre los peligros de que demasiada igualdad conduzca a una forma de tiranía de la mayoría. Por el contrario, Rousseau ve compatibilidad: su ideal de libertad política está basado en la igualdad moral y política de los ciudadanos y requiere de algún grado de igualdad socioeconómica. En la filosofía política contemporánea y en el pensamiento económico, libertarios como Nozick, Hayek, Berlin y Friedman siguen el punto de vista de la negociación y piensan que la libertad necesita ser protegida, mientras que pensadores como Habermas, Piketti y Sen creen que demasiada desigualdad hace peligrar la democracia, que necesita *tanto* de libertad como de igualdad (Giebler y Merkel, 2016). El siguiente capítulo se centra en los principios de la igualdad y la justicia y discute su relevancia para la política de la IA.

² En inglés, *freedom* y *liberty*. El autor afirma que utilizará ambos términos indistintamente. [N. del T.]

CAPÍTULO 3

Igualdad y justicia: sesgo y discriminación por la IA

INTRODUCCIÓN: SESGO Y DISCRIMINACIÓN COMO EJE CENTRAL PARA PLANTEAR PROBLEMAS RELACIONADOS CON LA IGUALDAD Y LA JUSTICIA

Las tecnologías digitales y los medios de comunicación no afectan solamente a la libertad, sino también a la igualdad y a la justicia. Van Dijk (2020) argumenta que la tecnología de redes, si bien contribuye a una producción y distribución más efectiva y eficiente, también conduce al aumento de desigualdades: «Globalmente, respalda una tendencia hacia el desarrollo combinado y desigual de los países [...]. Localmente, ayuda a crear economías duales con partes directamente enlazadas con la infraestructura de la información global y con partes que no lo son» (336). Esta divergencia en el desarrollo económico crea sociedades con diferentes «velocidades» de desarrollo. Algunas personas y países se benefician más que otros de estas tecnologías y medios de comunicación. Esta crítica se puede aplicar también a la IA. Y, como señalé en el capítulo anterior, la IA implementada en la robótica puede crear desempleo y, en consecuencia, más desigualdad. Los desarrollos hacia más automatización han venido progresando desde, al menos, el final del siglo XVIII. La IA facilita un paso más en esta revolución de la automatización, que crea beneficios para unos pocos (los propietarios de la tecnología de la IA y de los robots) y el riesgo de desempleo para la mayor parte. Este es un problema no solo para la libertad y la emancipación, sino también para la desigualdad. Como hemos visto, economistas como Stiglitz advierten contra una mayor desigualdad salarial y división social: la IA afecta a la sociedad en su conjunto. A menos que se tomen medidas para reducir estos efectos (por ejemplo, Piketty y sus colegas han propuesto impuestos más elevados por encima de un cierto

umbral —alto— [Piketty, Saez y Stantcheva, 2011], otros han propuesto una renta universal básica) mientras que el resultado sería una gran desigualdad y sus consecuentes problemas, como la pobreza.

Un problema que a menudo se encuentra relacionado específicamente con la IA y que plantea cuestiones relacionadas con la igualdad y la justicia es el sesgo. Como todas las tecnologías, la IA tiene consecuencias que los que la desarrollan no intentan producir. Una de ellas es que la IA, en la forma de aprendizaje automático, puede introducir, mantener y agravar sesgos y, por lo tanto, desventajas y discriminación para individuos o grupos concretos, por ejemplo para personas definidas en términos de raza o género. El sesgo puede surgir de varias formas: puede que haya sesgo en los datos de entrenamiento, en el algoritmo, en los datos a los que se aplica el algoritmo y en los equipos que programa la tecnología.

Un caso bien conocido es el del algoritmo COMPAS, un algoritmo de evaluación de riesgo para decidir sobre la libertad condicional, que fue usado en los Estados Unidos por el estado de Wisconsin: el programa de ordenador predice el riesgo de reincidencia (tendencia a reincidir). Un estudio (Larson *et al.*, 2016) descubrió que COMPAS atribuyó a los acusados negros un mayor riesgo de reincidencia del que realmente tenían, mientras que a los acusados blancos se les predecía un menor riesgo del que realmente tenían. El algoritmo, supuestamente entrenado con datos de decisiones pasadas, reprodujo el sesgo humano histórico e, incluso, lo aumentó. Además, Eubanks (2018) argumenta que la tecnología de la información como la IA y «el nuevo régimen de datos» tienen un impacto negativo sobre la igualdad económica y la justicia (8-9), ya que, a menudo, no benefician a los pobres ni a la gente de la clase trabajadora, ni tampoco los empodera, sino que les pone las cosas más complicadas. Las nuevas tecnologías son usadas para la manipulación, la vigilancia y el castigo de los pobres y desfavorecidos, por ejemplo en la forma de una toma de decisiones automatizada respecto de la elegibilidad para beneficios y sus consecuencias, conduciéndonos a un «asilo digital» (12). Mediante la toma de decisiones automatizada y el análisis predictivo de datos, los pobres son gestionados, moralizados e, incluso, castigados: «el asilo digital disuade a los pobres de acceder a los recursos públicos. Controla su trabajo, sus

gastos, su sexualidad y su parentesco. Intenta predecir su comportamiento futuro. Y castiga y criminaliza a aquellos que no se ajustan a sus dictados» (16). Además de socavar su libertad, sostiene Eubanks, esto crea y mantiene la desigualdad. Algunas personas (los pobres) son consideradas como si tuvieran un menor valor económico o político. A estos problemas se le añaden las desigualdades al acceso y uso de la información *online* en general (la denominada brecha digital). Por ejemplo, un menor acceso conduce a «menores oportunidades políticas, económicas y sociales» (Segev, 2010, 8), que pueden también enmarcarse como una cuestión de sesgo. El análisis de Eubanks (2018) muestra, además, que el uso de las tecnologías digitales está asociado con culturas específicas, en este caso una cultura de los Estados Unidos con «opiniones punitivas y moralistas sobre la pobreza» (16). El uso de la IA, una vez implementado por los Gobiernos, contribuye a la perpetuación de los sesgos.

Pero los problemas de desigualdad e injusticia con respecto a la IA y la ciencia de datos también ocurren fuera de las instituciones del Estado tales como el sistema jurídico, la Policía y la gestión del bienestar social. Considérese, por ejemplo, un banco que tiene que decidir sobre otorgar un préstamo: puede automatizar esta decisión subcontratándola a un algoritmo que calcula el riesgo financiero basándose en la historia financiera y laboral del solicitante, pero también sobre información como su código postal e información estadística sobre solicitantes previos. Si existe una correlación estadística entre tener un código postal concreto y no pagar préstamos bancarios, entonces alguien que viva en esa zona puede que no obtenga el préstamo, no basándose en la evaluación de su riesgo individual, sino sobre la base de patrones encontrados por el algoritmo. Si el riesgo individual es bajo, este caso parece injusto. Además, el algoritmo puede reproducir los sesgos inconscientes de los gestores bancarios que tomaron decisiones previas, por ejemplo un sesgo contra las personas de color. En respuesta al caso de la calificación automatizada para obtener un crédito, Benjamin (2019b, 182) advierte sobre la «“sociedad calificada” (*scored*)», en la que tener una u otra calificación crediticia es parte del diseño de la desigualdad». Aquellos que tengan peores calificaciones serán castigados. O, tomando un ejemplo (no estándar) del ámbito del género: ¿es justo que

un algoritmo decida que, ya que un conductor masculino joven tiene estadísticamente un riesgo mayor de accidente de coche basándose en una correlación entre género y accidentes, todos los conductores jóvenes masculinos tienen que pagar más por su seguro de coche, únicamente porque son varones, incluso si el riesgo puede ser bajo para un individuo en particular? Y, a veces, los conjuntos de datos están también incompletos. Por ejemplo, si un programa de IA está entrenado con conjuntos de datos que carecen de suficientes datos sobre mujeres, en concreto mujeres de color, mujeres discapacitadas y mujeres de la clase trabajadora, esto podría ser visto como un impactante caso de sesgo y desigualdad de género, como ha defendido Criado Perez (2019).

Incluso puede ser problemático para la mayoría de nosotros algo tan mundano como el uso de motores de búsqueda basados en IA. Noble (2018) argumenta que motores de búsqueda como Google refuerzan el racismo y el sexismo, y que esto debería considerarse como una «opresión algorítmica» (1), que se origina en las decisiones tomadas por seres humanos y controladas por las grandes corporaciones. Afirma que los algoritmos y los sistemas de clasificación se encuentran «incrustados internamente» (13) y afectan a las relaciones sociales, incluyendo las relaciones de poder raciales locales y globales. Señala que estas empresas ganan dinero con el racismo y el sexismo, y llama la atención sobre cómo los afroamericanos están afectados en términos de identidad, desigualdad e injusticia. Por ejemplo, el algoritmo de búsqueda de Google solía autoetiquetar a los afroamericanos como «simios» y relacionaba a Michelle Obama con el término «simio». Estos casos no son solamente insultantes u ofensivos; de acuerdo con Noble, también «demuestran cómo el racismo y el sexismo son parte de la arquitectura y del lenguaje de la tecnología» (9). La cuestión crucial no es que los programadores intenten codificar este sesgo; el problema es que ellos (y los usuarios de los algoritmos) dan por sentado que los algoritmos y los datos son neutrales, a pesar de que pueden estar incrustadas en ellos diferentes formas de sesgo. Noble advierte en contra de considerar a los procesos tecnológicos como si estuvieran descontextualizados y fueran apolíticos, una visión que encaja con la concepción de una sociedad en la que los individuos toman sus propias decisiones en el libre mercado (166).

Así pues, el problema no es simplemente que un algoritmo de IA en concreto esté sesgado en un caso concreto y tenga consecuencias concretas (por ejemplo, influencia política a través de los periodistas, quienes también usan motores de búsqueda como Google; véase Puschmann, 2018); el problema principal es más bien que estas tecnologías interactúan con, y apoyan a, las estructuras jerárquicas existentes en la sociedad y las concepciones e ideologías problemáticas que las alimentan. A pesar de que los usuarios no son conscientes de este hecho, estas tecnologías apoyan lógicas sociales, políticas y comerciales concretas y configuran el mundo de una forma concreta (Cotter y Reisdorf, 2020). Como los sistemas de clasificación *offline*, que reflejan los discursos más poderosos de la sociedad (Noble, 2018, 140), la IA nos puede conducir a la marginalización de las ideas y a la discriminación y opresión de las personas. Es más, a través de su alcance y velocidad, puede amplificarlas dramáticamente. Noble muestra que, como sucede con otras tecnologías digitales, la IA está «implicada» en la lucha por «la igualdad social, política y económica» (167) en un contexto de desigualdad e injusticia existente, y a veces creciente, y también en un contexto donde algunos discursos tienen más poder que otros y representan a personas oprimidas de formas específicas por aquellos que tienen más poder (141).

Debido a estas tensiones y luchas, los debates públicos sobre la IA en relación con el sesgo, la discriminación, el racismo, la justicia, la equidad, la discriminación de género, la (des)igualdad, la esclavitud, el colonialismo, la opresión, y así sucesivamente, normalmente se aprovechan, o se convierten enseguida, en debates altamente polarizados e ideológicos en contextos particulares (por ejemplo, los debates sobre el racismo en los Estados Unidos). Además, a pesar de que los informáticos y las empresas tecnológicas se han centrado en las definiciones técnicas del sesgo y la equidad, esto es necesario pero insuficiente para abordar todos los problemas sociotecnológicos (Startk, Greene y Hoffmann, 2021, 260-261). Como ya se señaló, investigadores como Noble, Eubanks y Benjamin han señalado correctamente el alcance más amplio que tienen estos problemas de sesgo y discriminación.

Sin embargo, como filósofos, debemos preguntarnos qué se quiere decir con los conceptos normativos usados en estas discusiones públicas, prácticas técnicas y literatura popular sobre el sesgo en la IA. Por ejemplo, debemos preguntar qué quiere decir justicia o igualdad, ya que esto conforma las respuestas a las preguntas respecto de los casos en los que está implicada la IA. ¿Hay algo incorrecto en este caso particular, y, si es incorrecto, *por qué* es incorrecto exactamente, y qué podemos y debemos hacer con ello y con qué *objetivo*? Para justificar nuestras opiniones, construir buenos argumentos y tener mejores discusiones sobre la IA sesgada, debemos (no simplemente los filósofos, sino también los ciudadanos, los que desarrollan la tecnología, los políticos, etc.) examinar los conceptos y los argumentos. Este capítulo muestra que algunas nociones y discusiones, en particular de la filosofía política, pueden ser muy útiles para este propósito.

Primero, proporcionaré una visión general de las discusiones filosófico-políticas estándar del ámbito anglosajón sobre la igualdad y la justicia para arrojar luz sobre lo que puede estar equivocado por lo que respecta al sesgo y la discriminación en términos de la IA. Plantearé qué tipo de igualdad y justicia están en juego aquí, y qué tipo de igualdad y justicia queremos. Pido al lector que considere diferentes concepciones de la igualdad y de la justicia. A continuación, volveré sobre dos reflexiones críticas sobre el pensamiento filosófico liberal sobre estas cuestiones. Los marxistas y los defensores de una política identitaria reclaman un cambio desde un pensamiento individualista, universalista, formal y abstracto hacia un pensamiento de clase, de grupo o basado en la identidad (por ejemplo, sobre la raza y el género). También prestan más atención a la discriminación en las vidas concretas de los desafortunados y al trasfondo histórico de esta discriminación: el colonialismo, la esclavitud, el patriarcado, el abuso, la historia de las relaciones sociales hegemónicas y el capitalismo. En ambos casos, mi propósito no es ofrecer una visión general de la filosofía política como tal, sino mostrar qué es lo que implica para la reflexión sobre el sesgo y la discriminación en, y en virtud de, la IA y la robótica.

¿POR QUÉ ESTÁ MAL EL SESGO (1)? IGUALDAD Y JUSTICIA EN LA FILOSOFÍA POLÍTICA LIBERAL ANGLOSAJONA ESTÁNDAR

Cuando se dice que la IA está sesgada, normalmente se hacen suposiciones no explícitas sobre por qué está sesgada y qué tiene esto de malo. Los filósofos pueden articular y discutir los argumentos. Un tipo de argumento se basa en la *igualdad*: si una recomendación o decisión basada en IA está sesgada, podríamos enmarcar este caso como una situación en la que se trata a las personas de forma desigual. Sin embargo, en la filosofía política hay importantes desacuerdos sobre qué implica la igualdad. Una manera de concebir la igualdad es la *igualdad de oportunidades*. En la concepción liberal universalista, «ciega», de la igualdad, se puede expresar del modo siguiente: las personas deberían tener una igualdad de oportunidades, sin importar su origen socioeconómico, su género, su origen étnico, etc.

¿Qué implica esto en el contexto de la IA? Imaginemos que un algoritmo de IA se usa para seleccionar candidatos para un puesto de trabajo. Dos criterios de éxito serán con bastante probabilidad la educación y la experiencia laboral relevante: los candidatos que puntúen más en estos ámbitos tendrán una mayor oportunidad de obtener del algoritmo la recomendación de ser contratados. Así pues, el algoritmo discrimina a las personas que tienen menor educación y menor experiencia laboral. Pero esto no se considerará normalmente como «discriminación» o «sesgo», ya que se supone que aquí la igualdad de oportunidades se respeta en el sentido de que todos los candidatos han tenido la oportunidad de adquirir el tipo de educación correcta y la experiencia laboral relevante para solicitar ese puesto de trabajo, sin importar criterios como el origen socioeconómico y el género. El algoritmo es «ciego» frente a estas características.

Sin embargo, algunos filósofos cuestionan esta concepción de la igualdad de oportunidades: argumentan que, en la práctica, algunas personas (por ejemplo, aquellas con un peor origen socioeconómico) han tenido unas menores oportunidades para adquirir un bagaje educativo y una experiencia relevantes. La igualdad de oportunidades real; de acuerdo con estos críticos, significaría que hemos creado las condiciones para que estas

personas menos favorecidas tengan unas oportunidades idénticas para conseguir la educación deseada y la experiencia laboral. Si esto no ocurre, el algoritmo discriminará en contra de ellas y su decisión podría denominarse sesgada debido a estas oportunidades desiguales. Si estos críticos se adhieren a una concepción liberal «ciega» de la igualdad, exigirán que todas las personas consigan oportunidades idénticas, sin importar de dónde vienen, qué pinta tienen, y así sucesivamente. Si la IA (a pesar de quizás las buenas intenciones) no ayuda a lograr esto, entonces está sesgada y hay que corregir este sesgo. Comenzando por una concepción no ciega (véase más abajo), se podría exigir entonces más oportunidades educativas y laborales para los desfavorecidos. Se podría también argumentar que, mientras este no sea el caso (mientras no se logre una igualdad de oportunidades ciega), el algoritmo tendría que decidir de una forma en la que se discrimina positivamente a favor de las personas con esa procedencia por lo que respecta a su clase social, género, etc. Estas son todas las diferentes justificaciones a favor de por qué la IA puede amenazar la igualdad, entendida en términos de oportunidades.

Estos argumentos ya señalan dos diferentes concepciones de la igualdad: una basada en la clase o en la identidad (véase la siguiente sección) y otra basada en la igualdad por lo que respecta a los resultados (en este caso, el empleo) en lugar de estarlo en las oportunidades. Las personas que quieren que el algoritmo discrimine positivamente a favor de estas clases o grupos específicos tienen un resultado particular en mente: una distribución particular en lo que respecta a los candidatos elegidos (por ejemplo, el 50 por 100 de aspirantes mujeres) y, en última instancia, una sociedad en la que los empleos estén distribuidos de una forma más equitativa y se acabe con las desigualdades históricas. La IA podría ayudar a alcanzar este resultado. Pero ¿qué implica la igualdad de resultados y cómo debería ser la distribución? ¿Implica que todos deberían tener lo mismo, que todos deberían tener un mínimo o que solo se deberían evitar las grandes desigualdades? Además, como pregunta Dworkin (2011, 347): ¿es la igualdad un valor en sí mismo?

En la filosofía política anglófona la igualdad no es un concepto muy popular. Muchas introducciones clásicas a la filosofía política ni siquiera

tienen capítulos sobre el tema (una excepción es Swift, 2019). Una forma más común de expresar el sesgo y por qué está mal radica en el concepto de justicia, en particular, de la *justicia como equidad* (Rawls, 1971; 2001) y de *justicia distributiva*. Se dice normalmente que el sesgo creado por el algoritmo no es *justo*. Pero ¿qué quiere decir justicia como equidad? Y si es que algo llega a necesitar ser redistribuido, ¿qué cuenta como una distribución justa? Aquí, también, existen diferentes concepciones. Considérese de nuevo el ejemplo de una IA que hace propuestas de empleo. Supongamos que tal IA, además de tener en cuenta criterios como la educación y la experiencia laboral, descubre que el código postal es una categoría estadísticamente relevante: imaginemos que existe una correlación entre el éxito para encontrar un trabajo y vivir en un barrio rico, «bueno» (socioeconómicamente hablando). El resultado podría ser que, siendo todo lo demás igual (por ejemplo, todos los candidatos tienen el mismo nivel de educación), un candidato de un barrio pobre y «malo» tendrá una menor oportunidad de ser elegido por el algoritmo. Esto parece injusto. Pero ¿qué, exactamente, es injusto y por qué?

Primero, se podría decir que es injusto porque, a pesar de que existe una correlación estadística, no existe una correlación causal: mientras que muchas personas que viven en ese barrio tienen, de hecho, una menor oportunidad de conseguir un empleo (debido a otros factores que están en juego como la falta de una buena educación), la persona en cuestión no tiene, ni debería tener, una menor oportunidad simplemente porque pertenezca a esta categoría estadística (tener el código postal X) teniendo además en cuenta que realmente tiene una buena educación y se desempeña bien en los demás indicadores. Esa persona es tratada de forma injusta porque la decisión está basada sobre un criterio que es irrelevante para este caso particular. Segundo, sin embargo, se podría preguntar también si es justo que muchas otras personas de ese barrio, que realmente han recibido una educación peor, una menor experiencia laboral relevante, etc., tengan una menor oportunidad de conseguir un empleo. ¿Por qué, como sociedad, permitimos diferencias tan grandes en este sentido? El problema podría, de nuevo, plantearse en términos de igualdad de oportunidades. Pero también podría formularse como una cuestión de justicia como equidad: la

distribución de la educación, la distribución de las oportunidades de obtener un empleo y la distribución real de los empleos parece injusta. La siguiente cuestión es, entonces: ¿por qué, exactamente, no son estas distribuciones justas, y cuál sería la distribución justa?

De acuerdo con una concepción de la justicia como equidad, una *igualitaria* y otra *redistributiva*, lo que se necesita es que todos obtengan lo mismo. En este caso significa lo siguiente: la política social y el algoritmo de IA deben asegurarse de que todos tienen las mismas oportunidades para obtener un empleo o que todos tengan un empleo (en cuyo caso no se necesitaría, para empezar, un algoritmo de selección). Mientras esta es una forma muy popular de habérmolas con la justicia distributiva en la mesa de la cocina o entre amigos (por ejemplo, cuando se necesita repartir una tarta), tendemos a sentirnos menos cómodos cuando se trata de política, empleo, etc. Muchas personas parecen creer que, cuando se trata de la sociedad, no es justa una distribución completamente igualitaria y que debería importar el mérito, que la gente talentosa merece más y (sorprendentemente, en mi opinión) que la riqueza y el apoyo heredados no suponen en absoluto ningún problema para la justicia. Nozick (1974), por ejemplo, piensa que la gente puede hacer lo que quiera con lo que poseen: están autorizados a hacerlo siempre y cuando lo hayan adquirido mediante transferencias voluntarias. Defiende un Estado mínimo que proteja los derechos a la vida, la libertad, la propiedad y el contrato, y rechaza el concepto de justicia redistributiva.

Pero, ya que el talento y el dinero heredado no se encuentran bajo el control de los individuos y son una cuestión de suerte, se podría argüir que esas circunstancias no deberían desempeñar ningún papel en la justicia y que las desigualdades basadas en ellas son injustas. Una concepción *meritocrática* de la justicia restringiría la selección exitosa a factores que tengan que ver con lo que la gente hace realmente, por ejemplo trabajar duro para conseguir un empleo. Un algoritmo justo sería entonces aquel que tuviera en cuenta el mérito. Sin embargo, esto también es problemático, ya que los criterios externos, como los grados o licenciaturas y otros resultados, no nos dicen, necesariamente, cuán duro ha trabajado una persona para obtener ese resultado. ¿Cómo sabemos qué ha hecho la

persona de nuestro ejemplo para conseguir el título? Por ejemplo, podría suceder que hubiese sido fácil para un solicitante obtener su título dado su trasfondo educativo o social. Y ¿cómo llegamos a conocer los méritos de la gente que vive en un barrio «malo»? Cuando consideramos su trasfondo y lo vemos en términos de malos resultados (en este caso, la carencia de un título), podemos pensar que no han hecho mucho por mejorar su posición, mientras que, en realidad, pueda no ser cierto en absoluto y que, por lo tanto, merezcan mucho más de lo que obtienen sobre la base de la llamada concepción meritocrática. La justicia entendida en términos de mérito puede ser justa, pero no es tan sencillo ponerla en práctica.

Sin embargo, incluso si se rechaza la justicia como igualdad distributiva absoluta, o la justicia basada en el mérito, existen otras concepciones de la justicia. Una afirma que es justo el que todos obtengan un mínimo de un bien particular (en este caso, la oportunidad de obtener un empleo). De acuerdo con la concepción *suficitaria* [*sufficitarian*] de la justicia (por ejemplo, Frankfurt, 2000; Nussbaum, 2000), necesitamos crear un umbral. En una sociedad tal, la gente que vive en barrios ricos aún seguiría teniendo una oportunidad mayor de ser elegida por el algoritmo. Pero la persona que vive en un barrio pobre tendría, independientemente de otros factores, un *mínimo* de oportunidades para conseguir el empleo. La correlación entre vivir en un barrio particular y obtener un empleo aún seguiría ahí, pero su relevancia en el proceso de decisión estaría debilitada. Este sería el caso, bien porque existe una política diferente que propicia este cambio antes o después de que el algoritmo haga su trabajo, o bien porque el algoritmo está ajustado de tal forma que otorga una mínima oportunidad a todo el mundo (un umbral de éxito), con otros factores que tienen el potencial de incrementar esa oportunidad, pero siendo imposible llegar a estar por debajo del umbral. De forma alternativa, todos pueden tener un mínimo de horas laborales (y, por tanto, ingresos) o un mínimo de dinero.

Sin embargo, de acuerdo con una concepción *prioritarista* [*prioritarian*] de la justicia, esto no sería aún justo. Las personas que viven en los barrios buenos aún tendrían unas oportunidades mucho mayores de conseguir empleos, y, cuando los tengan, sus trabajos serán a tiempo completo o estarán mucho mejor pagados. Lo que se necesita, de acuerdo con los

prioritaristas, es otorgar prioridad a los más desfavorecidos. En este caso, esto podría implicar una política que se centrara en dar empleo a las personas desfavorecidas (sin importar otros criterios) o en incrementar significativamente las oportunidades de conseguir un empleo para las personas que ya se encuentran en desventaja: por ejemplo, mediante un algoritmo que aumente las posibilidades de empleo para la gente que viva en un barrio pobre, incluso si obtienen menos puntuación en factores relevantes como la educación y la experiencia laboral.

Rawls propuso una famosa justificación filosófico-política para una posición prioritarista que responde también a la observación de que el talento es una cuestión de suerte y se construye en la igualdad de oportunidades. En su *Teoría de la Justicia* (1971), usó el experimento mental del denominado «velo de la ignorancia» en la «posición original» (12). Imagina que no sabes si nacerás con talento, si tendrás padres ricos o pobres, si tendrás igualdad de oportunidades, si vivirás en un barrio «bueno» o «malo», y así sucesivamente, y tampoco sabes qué posición social tendrás en la sociedad. En ese caso, ¿qué tipo de principios de justicia (y, por tanto, qué tipo de sociedad) escogerías? Rawls pensó que la gente propondría dos principios: uno que otorgue libertades iguales a todos, y otro que disponga las desigualdades sociales de tal forma que sean para el mayor beneficio de los menos favorecidos y creen igualdad de oportunidades. Las desigualdades están bien si maximizan la posición de los peor parados. A esto se le denomina el principio de la diferencia (60).

De acuerdo con estos principios de Rawls, el problema del algoritmo sesgado que selecciona sobre la base de códigos postales no reside en que su recomendación refleje una distribución desigual de los recursos socioeconómicos de la sociedad, o una sociedad en la que algunas personas caigan por debajo del umbral mínimo, sino que refleja y revela una sociedad en la que no hay igualdad de oportunidades y en la que las desigualdades no maximizan la posición de los peor parados. Si estos principios se hubieran implementado en las políticas efectivas, entonces probablemente no tendríamos una correlación tan alta entre el código postal y la oportunidad de conseguir un empleo. El resto de gente que vive en la zona habría tenido más oportunidades de conseguir un empleo y no estaría

en una posición social tan mala. Por lo tanto, el algoritmo habría encontrado solamente una correlación débil, y el código postal no habría desempeñado un papel tan importante en su recomendación. La situación de la persona que ha recibido una buena educación y con un buen bagaje, pero que vive en una zona pobre llena de gente con posiciones sociales que difieren enormemente de la suya propia no existiría, o el problema sería, como mínimo, menos pronunciado y, por tanto, este problema específico de discriminación por parte del algoritmo no habría surgido. Además, incluso si la situación actual es extremadamente injusta, se podría cambiar el algoritmo de tal modo que maximizase la posición de aquellos que se encuentran peor parados: una discriminación positiva que cambiase la situación actual de acuerdo con el principio de la diferencia de Rawls. Esto se podría denominar «discriminación positiva por diseño», como una forma específica de «equidad por diseño».

Obsérvese que esta forma de discriminación positiva necesitaría, para empezar, que los programadores y diseñadores sean conscientes del sesgo potencial, especialmente del sesgo *no intencionado*. De forma más general, tendrían que ser conscientes de que incluso si no se pretende producir discriminación u otras consecuencias políticas relevantes, las elecciones del diseño pueden conllevar tales consecuencias, por ejemplo en términos de justicia e igualdad. Aún hay mucho trabajo que hacer para crear consciencia de las consecuencias políticas potenciales, para reconocer el sesgo y, de forma más general, para implementar valores políticos y éticos en el diseño. Por ejemplo, reconocer el sesgo puede ser difícil cuando no hay referencias explícitas a criterios tales como el género, la raza, etc., en los datos de entrenamiento (Djeffal, 2019, 269). Y cuando el problema no se reconoce, no puede haber una solución, incluso en términos de discriminación positiva. El trabajo técnico sobre equidad algorítmica puede ayudar en esta tarea: intenta identificar, medir y mejorar la equidad cuando se usan algoritmos de IA (por ejemplo, Pessach y Shmueli, 2020). Combinado con el tipo correcto de marco legal, esto puede conducir a lo que Hacker (2018, 35) llama «tratamiento igualitario por diseño». Pero, como hemos visto, la igualdad es solamente una forma de encuadrar los problemas. Además, el diseño se puede utilizar también para discriminar positivamente. En tal

caso, el objetivo y la definición de equidad algorítmica no consiste en que los resultados sean independientes de variables como el género, el origen étnico, etc., para evitar un sesgo negativo, sino, por el contrario, en que se cree un sesgo positivo hacia una o más de tales variables, corrigiendo así la injusticia histórica o actualmente existente.

Sin embargo, como veremos ahora mismo, las medidas de discriminación positiva se proponen normalmente no por aquellos que trabajan en la tradición filosófico-liberal, sino por aquellos que critican esa tradición o, al menos, critican su universalismo.

¿POR QUÉ ESTÁ MAL EL SESGO (2)? LAS TEORÍAS DE LA IDENTIDAD Y DE CLASE COMO CRÍTICA DEL PENSAMIENTO LIBERAL UNIVERSALISTA

La teoría marxista ha criticado las explicaciones filosófico-liberales de la justicia y la igualdad por centrarse en principios formales y abstractos sin tocar la estructura capitalista básica de la sociedad, de acuerdo con la cual los individuos formalmente libres realizan contratos voluntariamente (véase, de nuevo, a Nozick), pero que realmente crea y mantiene una división y jerarquía entre dos clases: una que posee los medios de producción y otra que es explotada por la anterior bajo condiciones capitalistas. En lugar de imaginar posiciones y contratos hipotéticos, deberíamos posar la vista sobre las condiciones materiales e históricas que crean la injusticia y la desigualdad y cambiarlas. En lugar de mantener separadas las cuestiones relacionadas con la producción y la distribución, deberíamos cambiar la forma en que organizamos la producción. En este sentido, una sociedad comunista se encontraría más allá de la justicia (Nielsen, 1989), al menos, si la justicia se entiende como justicia redistributiva. En lugar de, como mantiene la teoría liberal, tener primero una producción capitalista y luego redistribuirla de acuerdo con principios de justicia, deberíamos abolir el capitalismo mismo. En lugar de evaluar la sociedad desde un punto de vista desinteresado, deberíamos defender los intereses de la clase que es explotada. En lugar de hablar sobre principios de justicia aplicados a individuos y agregados de individuos, deberíamos centrarnos en la clase y en la lucha de clases.

Así pues, desde esta perspectiva los algoritmos sesgados y las sociedades a las que se aplican no son injustas, no equitativas, etc., porque fallen a la hora de aplicar y de encarnar una idea abstracta de justicia o igualdad, sino porque ayudan a crear y a mantener un sistema socioeconómico, el capitalismo, que crea relaciones sociales jerárquicas entre dos clases: la de aquellos que poseen y la de aquellos que no poseen los medios de producción. Los problemas planteados por la teoría liberal se encuentran enmarcados dentro de un mundo capitalista. Considérese de nuevo el ejemplo del préstamo o el de la contratación: ambos casos tienen lugar en una estructura social y económica capitalista en la que está en el interés de los capitalistas mantener una clase de personas con deudas y en unas posiciones socioeconómicas precarias en disposición de ser explotadas. En consecuencia, este sesgo no se encuentra solamente en el algoritmo o en una situación social particular. Dentro del capitalismo mismo existen un sesgo y una dinámica que favorecen a unos (los capitalistas como poseedores de los medios de producción) y desfavorecen a otros (la clase trabajadora, que se convierte en el proletariado). La IA se usa como un medio de explotación y los robots se usan para reemplazar a los trabajadores y crear un proletariado de personas desempleadas, que hace más sencillo explotar a aquellos que aún tienen trabajo. El problema no es la IA, sino lo que denominaríamos «capitalismo de IA» (para usar un término más general que el de Zuboff [2019], «capitalismo de la vigilancia», y para subrayar el papel de la IA). A menos que se aborde este problema fundamental, no habrá ni justicia ni igualdad. Se podrían corregir los algoritmos para beneficiar a los desfavorecidos, pero, en última instancia, estos son tratamientos meramente sintomáticos. El problema real es que la IA y la robótica se usan dentro de un sistema capitalista, que utiliza estas tecnologías no para la emancipación de la gente, sino por el único beneficio de hacer a los capitalistas aún más ricos de lo que ya son. Además, la creencia de que, con el tiempo, el libre mercado eliminará la discriminación algorítmica es «infundada», ya que aquellos que emplean los algoritmos no tienen incentivos para minimizar el sesgo (Hacker, 2018, 7). En el ejemplo, el banco y la empresa que contratan empleados operan dentro de una lógica capitalista y no creen que sea su trabajo reducir la

discriminación. Si esto no se cambia, los tratamientos sintomáticos no ayudarán demasiado. Ni habrá siquiera un incentivo por parte de los capitalistas para cambiar realmente las tecnologías y su uso, ya que no les conviene.

Visto desde esta perspectiva, es entonces importante para los trabajadores y para otras personas oponer resistencia al sistema y luchar contra la IA impulsada por el capitalismo. Sin embargo, uno de los problemas es que, a menudo, estos no saben que se está usando la IA, y no digamos ya qué es lo que categoriza y discrimina. Su funcionamiento y contribución a la hora de crear sesgos se encuentran ocultos. Además, el impacto del capitalismo de IA sobre los trabajadores no es idéntico en todos los casos. Algunos trabajos son más precarios que otros, aunque, hasta cierto punto, todos los trabajos se han vuelto más precarios. Azmanova (2020, 105) habla de «capitalismo de la precariedad», afirmando que «la inseguridad económica y social se ha convertido en una característica fundamental de nuestras sociedades», lo que conduce a la ansiedad y al estrés (Moore, 2018). Incluso aquellos que tienen puestos de trabajo cualificados y bien remunerados no están seguros. Sin embargo, algunos trabajos son claramente más precarios que otros, y algunos trabajadores y algunos egos reciben una recompensa mayor que otros (ver capítulo 5). Esto también implica que las consecuencias psicológicas del capitalismo contemporáneo están distribuidas desigualmente: algunos conforman «egos ansiosos que han internalizado el imperativo de rendir» (Moore, 2018, 21) más que otros. Algunos temen el reemplazo por las máquinas (15) más que otros. Y, mientras que los trabajadores de bajo estatus se encuentran altamente expuestos a la vigilancia sin posibilidad de evitarla, los trabajadores de mayor estatus están más protegidos, a pesar de que sus datos son también explotados (Couldry y Mejias, 2019, 191). Todos son vulnerables existencial, socioeconómica y psicológicamente, pero algunos son más vulnerables que otros. Además, existen diferencias culturales en cómo se percibe la IA: algunas culturas tienen actitudes más positivas que otras hacia ella (y hacia la tecnología en general). (Esto también afecta a los desafíos relacionados con la regulación de la IA, especialmente a nivel global. Volveré a este punto en mi conclusión). En conjunto, esto implica

que algunos estarán más motivados a rebelarse contra el capitalismo de la IA que otros, si es que son conscientes en uno u otro modo de los problemas relacionados con la tecnología. Esto cuestiona el ideal marxista de una amplia alianza entre trabajadores bajo el paraguas de una única clase (conciencia).

Con todo, el cambio social no incide solamente sobre la gente, sus acciones y el trabajo. La tecnología, incluyendo la IA, es una parte esencial del sistema. En su libro *Inhuman Power* (2019), Dyer-Witheford, Kjøsen y Steinhoff (algunos de los más destacados analistas de la tecnología de inspiración marxista) argumentan que la IA debería verse como la culminación de la alienación de los trabajadores bajo el capitalismo: la IA representa el poder del capital autónomo, que conduce a la mercantilización y a la explotación. Esto se puede formular como un problema de libertad (ver el capítulo previo), pero también como un problema del capitalismo, que crea una profunda desigualdad e injusticia. También puede ser un problema para otros valores políticos. Por ejemplo, como defiende Frankfurt (2015), la desigualdad económica es también un problema para la democracia: «aquellos que se encuentran en mejores circunstancias tienen una importante ventaja sobre aquellos que están menos acomodados: una ventaja que pueden tender a explotar para conseguir una influencia inapropiada sobre los procesos electorales y regulatorios» (6).

Liberales como Frankfurt defienden que la desigualdad «no puede ser en sí misma nuestra mayor ambición esencial» (5). Sin embargo, para los marxistas, el problema no son los efectos sobre la democracia, sino la desigualdad misma, en tanto que conectada con la explotación capitalista. Además, mientras que los pensadores liberales (de izquierda) procederían a exigir una redistribución sin decir nada sobre los medios de producción (en este caso, IA), para los marxistas este es un punto clave. Por ejemplo, Dyer-Witheford sugiere que el cambio social necesitaría que cambiase el sistema socioeconómico al completo, así como la tecnología, ya que la IA y el capital se encuentran demasiado imbricados entre sí. De forma paradójica, si el proletariado quiere adoptar medidas contra el capital, deberá usar la IA, pero, al mismo tiempo, estar en contra de ella (Dyer-Witheford, 2015, 167). Quizás se podría comprender este punto bajo la luz de lo que escribió

Marcuse (2002) sobre la opinión de Marx de que es necesaria una reorganización de los medios de producción. De acuerdo con Marx, la producción debería estar organizada por los productores inmediatos. Pero, cuando la tecnología «se convierte en el medio de control y cohesión en un universo político que incorpora a las clases trabajadoras», argumenta Marcuse, necesitamos «un cambio en la propia estructura tecnológica» (25).

Esta idea sugiere que la propia tecnología necesita cambiarse. Sin embargo, la mayoría de los marxistas se centran en la *propiedad* de los medios de producción, sin cuestionar a las tecnologías mismas. De acuerdo con Fuchs (2020), una sociedad verdaderamente justa debe estar basada en bienes comunes, lo que significa que la información debería ser un bien común en lugar de una mercancía, y que debería haber un control común sobre las condiciones de comunicación. El capital intenta subsumir los bienes comunes. Fuchs propone, al contrario, que los trabajadores deberían «obtener el control colectivo de los medios de comunicación como medios de producción económica» (310) y que plataformas tales como Facebook deberían ser cooperativas basadas en la sociedad civil (311). Fuchs describe este «antagonismo entre información como un bien común y como mercancía» en términos de justicia e igualdad:

Si la forma de la mercancía implica desigualdad, entonces una sociedad realmente equitativa, democrática y justa debe ser una sociedad basada en los bienes comunes. Para el sistema de comunicación, esto significa que los sistemas de comunicación como bienes comunes corresponden a la esencia de la humanidad, de la sociedad y de la democracia (28).

De forma similar, se podría argumentar que la IA y los datos, entendidos como tecnologías de la comunicación y de la información, son medios de producción en un sentido marxista y deberían mantenerse como bienes comunes en lugar de estar controlados por el capital. Además, se podría cuestionar el optimismo y el determinismo tecnológicos de la visión transhumanista de la IA (Fuchs la denomina «poshumanista»), que parecen suponer «que la sociedad y la humanidad cambian radicalmente debido al surgimiento de nuevas tecnologías» (21) y que este cambio es necesariamente bueno. Fuchs advierte que esta idea olvida la importancia de la clase y del capitalismo en la sociedad (21), conduce a la sustitución de las personas por robots y a una concentración del poder, en lugar de

llevarnos a la democracia y a la igualdad (82). Además, se ha argumentado que pensar que las nuevas tecnologías de la información y de la comunicación son necesariamente progresivas implica negar las «condiciones antagonistas» de su surgimiento y «su integración dentro de las brutalidades del capitalismo global» (Dean, 2009, 41).

Un enfoque alternativo, actualmente muy popular en los Estados Unidos, que es crítico con el enfoque filosófico liberal clásico de la justicia y de la igualdad, se centra no en las categorías socioeconómicas, sino en las categorías que tienen que ver con la identidad, como la raza y el género. A veces se denomina a este enfoque «política identitaria». La frase misma está muy cargada políticamente y es controvertida; se refiere a «un amplio abanico de actividad y teoría política que se encuentra en las experiencias de injusticia compartidas por los miembros de ciertos grupos sociales» (Heyes, 2020). Si la política identitaria usa principios políticos como la libertad, la justicia y la igualdad, es para asegurar estos principios para grupos específicos, definidos en términos de su identidad e historia. Mientras que los teóricos de la tradición filosófica liberal toman una posición universalista (por ejemplo, exigen justicia para todos o igualdad para todos), las personas que abrazan el pensamiento de la política identitaria argumentan que el pensamiento de tal tradición no ha sido suficiente para detener la marginalización u opresión de grupos sociales concretos como las mujeres, las personas de color, las personas LGBTQ+, los indígenas, las personas discapacitadas, y así sucesivamente. Para abordar estos problemas, colocan a la identidad (de grupo) en el centro de la atención política. Se podría decir que levantan el velo de ignorancia de Rawls. En lugar de participar en un experimento mental distante sobre individuos abstractos y sociedades que están compuestas por estos individuos, nos piden que nos fijemos en realidades concretas e historias de injusticia contra (personas de) grupos concretos. Como los marxistas, desean el cambio de las estructuras sociales que generan injusticia y buscan en realidades históricas concretas en lugar de apelar a nociones universalistas abstractas. Sin embargo, no hacen estas exigencias tanto porque una clase socioeconómica se encuentre desfavorecida o porque el capitalismo sea problemático, sino porque grupos concretos definidos

identitariamente se encuentran actual e históricamente desfavorecidos. Además, si se empareja con una política de la diferencia, se exige justicia, pero con el objetivo no de una inclusión en la humanidad universal *independientemente* de las identidades y las diferencias, sino para respetar estas propias identidades y diferencias. Esto también implica reconocer grupos y personas particulares como pertenecientes a estos grupos, en lugar de pensar sobre un «nosotros» universal. Fukuyama (2006) explica así esta idea: desde Hegel, la política se ha conectado con el reconocimiento, pero ahora

el reconocimiento universal basado en una humanidad compartida no es suficiente, en concreto por parte de los grupos que han sido discriminados en el pasado. De ahí que la política identitaria moderna gire en torno a las exigencias de reconocimiento de identidades de grupo (9).

Estas identidades se encuentran históricamente situadas y localmente construidas, y a menudo surgen contra formas particulares de opresión e injusticia.

Hoy día, esta forma de política es popular entre la izquierda liberal. Lo que Dyer-Witheford llama la posición «posmarxista» rechaza la teoría marxista como totalizadora y reductiva, afirmando que es ciega ante el patriarcado y el racismo y que niega la diversidad cultural. En su lugar, los posmarxistas prestan atención a la diferencia, al discurso y a la identidad, y hablan no sobre revolución, sino sobre democracia (Dyer-Witheford, 1999, 13). En parte, esta postura puede considerarse como una continuación de la política posmoderna. En lugar de argumentar en contra del capitalismo y ofrecer una visión de la solidaridad, muchos han comenzado a subrayar la diferencia y la identidad. Esta postura no desafía realmente al capitalismo, sino que más bien coexiste a menudo de manera bastante fácil con él, por ejemplo en la moda (Dean, 2009, 34). De forma similar, el énfasis posmoderno en las identidades fluidas y altamente individuales es fuertemente compatible con la ideología neoliberal, pero parcialmente es más que eso: existe una conciencia de las formas históricas de injusticia y se ha adoptado de nuevo una parte de la retórica marxista de la lucha, la resistencia y el cambio sistémico, si bien sin poner el foco en las categorías de clase y socioeconómicas, y con un rechazo al universalismo.

Con respecto al sesgo en la IA y a la discriminación por medio de la IA, la cuestión normativa relevante es, entonces: ¿contra qué grupos específicos ejercen una discriminación estas tecnologías y quienes las emplean, y qué luchas por el reconocimiento de ciertas injusticias históricas están amenazadas por la IA? ¿Existe un sesgo contra las mujeres? ¿Contra las personas transgénero? ¿Contra las personas de color? ¿Contra los discapacitados? Un argumento famoso en el área de la IA y de la robótica que hace uso de la raza es el de Benjamin (2019a; 2019b), que ha argumentado que estas tecnologías no son políticamente neutrales, sino que agudizan la discriminación racial, la desigualdad y la injusticia. Escribiendo en el contexto de una historia (y, por desgracia, presente a menudo) de la discriminación racial en los Estados Unidos, Benjamin afirma que la injusticia se lleva a cabo contra un grupo particular, las personas de color. Desde esta perspectiva, lo que está mal en los algoritmos sesgados en cuestión es que no son neutrales, sino que desfavorecen sistemáticamente a los negros, reproduciendo así las desigualdades existentes y contribuyendo a «formas entrelazadas de discriminación» (Benjamin, 2019b), en concreto el sesgo racial. Rhee (2018, 105), mientras tanto, ha afirmado que la aparición de muchos robots de compañía y muñecos inteligentes «normalizan lo blanco». Tales argumentos difuminan la visión instrumentalista de que la tecnología es neutral y se oponen radicalmente a las narrativas (a menudo provenientes del mundo de las grandes corporaciones) sobre la neutralidad de la tecnología y sobre lo digital como lugar en el que hay un «terreno de juego equilibrado» e, incluso, «donde las desigualdades están rectificadas» (Benjamin, 2019b, 133). En este caso, se miran los problemas a través de la lente de la raza y de la identidad, en lugar de hacerlo desde un punto de vista desconectado y universalista.

Considérese de nuevo el caso del arresto injustificado del que hablamos al comienzo de este libro. De acuerdo con Benjamin y otros que adoptan una perspectiva similar sobre la identidad, lo que está equivocado en este caso (y en otros similares) no era que una «persona» o un «ciudadano» en general y en abstracto fuese detenido injustamente, sino que un hombre negro fue arrestado *porque era negro*: esto era racismo. O, por ponerlo en términos de un eslogan popular usado en el contexto de una década de

protestas en los Estados Unidos contra la violencia racialmente motivada: el enfoque de este argumento basado en la identidad no es que «todas las vidas importan» (*all lives matter*), sino que «las vidas negras importan» (*black lives matter*). En lugar de lo que perciben como la sublime perspectiva «blanca» de la teoría liberal clásica, pensadores como Benjamin prefieren mirar a través de la lente de la raza lo que está ocurriendo realmente al nivel del suelo. Señalan que el pensamiento universalista no ha sido efectivo a la hora de crear una sociedad justa e igualitaria y afirman que solo ha satisfecho a grupos específicos (por ejemplo, las personas blancas, los hombres). Por ejemplo, en lugar de apelar a principios universalistas, Benjamin (2019b) reclama imaginar «alternativas al *techno quo* (la situación normal cuando se trata de tecnociencia)» mediante una «imaginación liberadora» que se «desarrolle sobre una tradición racial negra» (12). En este caso, la imaginación política se alimenta de historias concretas (de grupo o de identidad), en lugar de apelar a concepciones abstractas de justicia o igualdad.

Un aspecto crucial del pensamiento identitario es la referencia al trasfondo histórico. Por ejemplo, cuando se trata de identidad racial, los defensores de la política identitaria señalan los horrores históricos de la esclavitud y el colonialismo. Esto puede hacerse de, al menos, dos formas compatibles. Una es argumentar que las injusticias *presentes* realizadas contra, por ejemplo, las personas negras en los Estados Unidos, no solamente tienen raíces en el racismo (como si esto fuera solo una especie de sistema abstracto de creencias), sino que son también la continuación de estas formas históricas de opresión y prácticas racistas concretas terriblemente equivocadas, si bien en formas que no son formalmente reconocidas como esclavitud y colonialismo. Reparar el sesgo en la IA es entonces una forma de contribuir a la lucha contra el racismo y contra la abolición total (no parcial) de estas formas de opresión, evitando su continuación en el futuro. Las críticas que se centran en el (neo)colonialismo también adoptan esta perspectiva histórica. Por ejemplo, Couldry y Mejias (2019) hablan del «colonialismo de datos» para expresar cómo las desigualdades presentes son una continuación de formas históricas de imperialismo y explotación, al menos «hasta el punto en el que estamos

siendo explotados a través del uso de los datos que creamos» (107-108). Los datos de las personas, el trabajo y, en última instancia, las relaciones sociales (12) son apropiados por el capitalismo. Además, el uso de la IA en un lugar (por los privilegiados) se basa en el trabajo y la explotación en otra parte, muy lejos de allí. Por ejemplo, el entrenamiento del aprendizaje automático puede implicar tal forma de explotación.

La perspectiva histórica también nos conduce hasta el tema del colonialismo. Teniendo presente el trasfondo del colonialismo histórico, se puede mirar críticamente a la IA actual y a otras prácticas técnicas refiriéndose al peligro (¿o realidad?) del neocolonialismo. Por ejemplo, existe la preocupación de que los humanos e individuos universales a los que se dirige la teoría liberal son, en la práctica, personas que viven en sociedades occidentales acomodadas y que los intereses y la identidad del hemisferio sur se encuentran, por lo tanto, olvidadas. Mientras que esta preocupación se puede también formular dentro del liberalismo y del marxismo, por ejemplo refiriéndose a las desigualdades e injusticias socioeconómicas, o a la opresión de los trabajadores y a la geopolítica dentro de un contexto capitalista, también se puede plantear en términos de identidad y colonialismo. Por ejemplo, un artículo de opinión habla sobre el «colonialismo digital» como la explotación imperialista de los países más pobres por las poderosas corporaciones tecnológicas de Silicon Valley (Kwet, 2019). De forma similar, se ha discutido (Birhane, 2020) que la «invasión de África por parte de la IA repite la era colonial de la explotación» con el abandono de las necesidades e intereses locales y la perpetuación del sesgo histórico de una forma en la que se desfavorece a los grupos minoritarios (por ejemplo, aquellos que no tienen documentos, que están excluidos del sistema nacional biométrico), equivale a la «colonización algorítmica» de ese continente. (Algunos de estos problemas también pueden abordarse en términos de transferencia de tecnología, que a menudo perpetúa las prácticas no democráticas en países en desarrollo. Este hecho se puede considerar como una forma de colonialismo e injusticia contra grupos específicos, pero también se puede formular en términos de violaciones de los derechos humanos.)

Otra forma compatible de incorporar la historia es usarla para alertar sobre las injusticias presentes, que puede que sean menos malas que las históricas, pero que pueden conducir a formas peores de opresión en el *futuro*, y a las que se tiene que poner freno antes de que se alcance ese punto. En este ejemplo se considera que el racismo y otras formas actuales de opresión conducen necesariamente a sociedades que sistemáticamente oprimen y explotan a las personas de un origen racial concreto; en otras palabras: en sociedades basadas en el pensamiento colonial y esclavista. Esta es la razón por la que se tiene que poner coto al racismo en y a través de la IA. Otro ejemplo es el género: el sesgo de género en la IA es considerado como la continuación de formas históricas de opresión y patriarcado y como un problema que puede conducir a nuevas formas de opresión y patriarcado. Por ejemplo, en internet puede haber sesgo en el lenguaje y en cuerpos lingüísticos concretos (por ejemplo, un sesgo que conecta una profesión particular como la de doctor en la medicina con los hombres), y si este sesgo alimenta a las herramientas de inteligencia artificial y de la ciencia de datos para el procesamiento de lenguajes naturales (Caliskan, Bryson y Narayanan, 2017; Sun *et al.*, 2019), entonces perpetuará el sesgo histórico que se encuentra presente en los textos y potencialmente agravará el sesgo en el futuro.

Sin embargo, el centro de atención a menudo se coloca sobre la continuación del sesgo histórico. Por ejemplo, haciendo un comentario sobre la policía predictiva, Crawford y Calo (2016) usan el lenguaje de la política identitaria cuando dicen en *Nature* que necesitamos investigar cómo «los sistemas de IA afectan desproporcionadamente a grupos que ya se encuentran desfavorecidos por factores como la raza, el género y el trasfondo socioeconómico» (312). A menudo se afirma también que la IA puede reflejar un sesgo contra las mujeres debido al sesgo presente en los equipos de los que desarrollan la tecnología, predominantemente compuestos por hombres, quienes, a su vez, heredan el sesgo de generaciones previas. Así, el punto de referencia es tanto una forma histórica como presente de discriminación contra grupos específicos, que están caracterizados en términos de identidad (raza, género, etc.) en lugar de, por ejemplo, solo criterios socioeconómicos. Los casos individuales de

discriminación y opresión aquí y ahora se analizan a la luz de la discriminación histórica contra grupos específicos definidos en términos de identidad y buscando el reconocimiento de esa identidad y de esa historia. La perspectiva histórica es compartida por el marxismo, pero el punto de atención se desplaza desde la defensa de los intereses de una clase socioeconómica particular y la promoción de la meta de la emancipación universal (marxismo) hasta las preocupaciones presentes, la historia y el futuro de ciertos grupos particulares definidos por su identidad.

También se pueden presentar argumentos similares basados en la historia contra el uso y la comprensión de robots como «esclavos». El argumento *a favor* de usar robots como esclavos es que frena la explotación de los humanos. En el contexto de una argumentación en contra de otorgar derechos a los robots, Bryson (2010) ha afirmado que los robots deberían ser considerados legalmente como esclavos. Se podría, entonces, lidiar con cuestiones legales usando el derecho romano, de acuerdo con el cual el propietario de los esclavos es responsable de los daños y perjuicios que pudieran causar, como ha propuesto Floridi (2017). Sin embargo, como he sugerido en el capítulo anterior, algo parece ir mal cuando pensamos sobre los robots como esclavos. La perspectiva de la política identitaria proporciona ahora un argumento de por qué esto es erróneo, un argumento que se puede utilizar en lugar de, o además de, la objeción marxista. Vista la cuestión con el trasfondo histórico de la esclavitud y de la discriminación contra grupos específicos, se podría objetar la consideración de los robots como esclavos argumentando que perpetúa una fase de la historia del pensamiento en términos de amo y esclavo y una historia de marginalización y exclusión de un grupo específico. A pesar de que no se daña a los humanos cuando los robots se usan como esclavos, sin embargo, el tipo de discurso y reflexión sobre las relaciones sociales se considera como fundamentalmente problemático. Así, la objeción en contra de considerar a los robots como esclavos puede recibir apoyo por parte del pensamiento liberal universalista o marxista, por ejemplo mediante argumentos contra las relaciones sociales hegemónicas y capitalistas, o mediante argumentos identitarios que relacionan la esclavización de los robots con la marginalización histórica, presente y quizás futura de grupos

específicos (de humanos). Hoy en día, los robots son los esclavos, pero ¿quizás mañana esto se podría extender a otros grupos? Además, se podría extender el alcance de este argumento a los animales: ¿no es una forma de esclavitud nuestro tratamiento de algunos animales no humanos? Diré algo más sobre la política con respecto a los no humanos en el capítulo 6, pero aquí es interesante apuntar que todas estas críticas comparten la opinión de que ni el pensamiento griego ni el romano sobre los esclavos, ni desarrollos posteriores de la historia de la filosofía que suponen o apoyan estos tipos de jerarquías sociales son buenas fuentes para la evaluación normativa de las tecnologías, ya que continúan una forma de pensamiento hegemónico y colonial. Es, por lo tanto, importante estudiar críticamente los discursos de exclusión y dominación política que puedan acompañar el uso y desarrollo de la IA.

El pensamiento identitario continúa siendo controvertido, en la filosofía política y en todas partes. Incluso en la teoría feminista, que a menudo simpatiza con el pensamiento identitario, ha habido discusiones sobre lo que implica la identidad femenina, por ejemplo si es una identidad esencial o una identidad que tenga que ser entendida en términos *performativos* (Butler, 1999). Las feministas contemporáneas han reclamado una política posidentitaria que vaya más allá de «una política parroquial del reconocimiento» y de una política del sufrimiento basada en la identidad de las víctimas y que, en cambio, preste atención a «una política de la diversidad más amplia» (McNay, 2010, 513). Por ejemplo, McNay (2008) ha argumentado en contra de reducir la política a cuestiones de identidad. Pero, como era de esperar, ha habido también críticas tanto desde los liberales como desde los marxistas. Los marxistas acusan a los defensores de la política identitaria de ocuparse solamente de la superestructura (la cultura) sin prestar atención a la economía subyacente, centrándose así en grupos marginados, pero descuidando las estructuras económicas en general que producen desigualdad y continúan atadas a la lógica capitalista. Prefieren usar la categoría de «clase» para analizar y desafiar la desigualdad socioeconómica bajo el capitalismo. Y Fukuyama (2018a) advierte de que las sociedades guiadas por la identidad hacen que la deliberación y la acción colectivas sean complicadas, ya que se fracturan en segmentos de

identidad. También señala que la derecha usa el mismo lenguaje, por ejemplo para victimizar a los hombres blancos, mientras que la izquierda se divide en «una lista de grupos de identidad» (Fukuyama, 2018b, 167). Propone que en tales democracias necesitamos «buscar nuestro camino hasta una comprensión más universal de la dignidad humana» (xvi).

Para reflexionar sobre el sesgo y la IA, estas discusiones y sus contextos políticos (por ejemplo, el movimiento *Black Lives Matter*) siguen siendo relevantes. Cómo plantear el sesgo y la discriminación importa mucho no solo filosóficamente, sino también en la práctica. ¿Deberíamos desarrollar principios universalistas de justicia en nuestros algoritmos o centrarnos, en cambio, en el sesgo respecto de grupos específicos y tomar medidas de discriminación positiva a través de la tecnología y de otras maneras? ¿Deberíamos aceptar el sesgo histórico presente en los cuerpos lingüísticos (por ejemplo, afirmar que los datos importados de internet reflejan la sociedad tal cual es, que los datos y algoritmos deben ser «neutrales») o insistir en que los algoritmos nunca son neutrales, que están sesgados, y corregir este sesgo, contribuyendo de forma efectiva a una sociedad que esté menos sesgada respecto de grupos históricamente desfavorecidos? ¿Podrían combinarse ambos tipos de argumentos o son inconmensurables? ¿Podrían ser implementados en la práctica aspectos de ambas teorías normativas? Si es así, ¿qué tensiones potenciales pueden surgir?

CONCLUSIÓN: LA IA NO ES POLÍTICAMENTE NEUTRAL

Desde la filosofía de la tecnología sabemos que las tecnologías no son, ni pueden ser, moral y políticamente neutrales. Es así en general (para todas las tecnologías), y es también el caso para la IA y la ciencia de datos. A pesar de que algunas personas piensan que, como señala Matzner (2019, 109), «los algoritmos podrían ser neutrales si los humanos no los estropearan constantemente con sus sesgos», esta visión está equivocada. En efecto, la relación entre los humanos y las máquinas es bastante más compleja, y esto es también cierto para los humanos y para la IA. Los algoritmos de IA nunca son neutrales y tanto el sesgo en la sociedad como los sesgos que resultan de los algoritmos y los procesos de la ciencia de

datos necesitan ser evaluados. De forma similar, como han discutido Gitelman y Jackson (2013), los datos mismos no son neutrales, objetivos ni «crudos». Por el contrario, están «producidos por operaciones de generación de conocimiento» (3), que implican interpretación, gestión y percepción incluyendo incluso sentimientos (Kennedy, Steedman y Jones, 2020). Como ya he mencionado, podría, por ejemplo, haber sesgo en los cuerpos lingüísticos usados por la IA. El sesgo se encuentra también integrado en el lenguaje mismo, por ejemplo el sesgo de género. Considérese el uso de la palabra «hombre» de forma que se supone que se refiere no solo a los varones, sino también a la especie humana en general (Criado Perez, 2019, 4). Además, como ya he sugerido, los equipos que tratan con los datos tampoco son neutrales. Las personas pueden tener sesgos y los equipos pueden carecer de diversidad étnica y de opinión política, por ejemplo cuando están formados principalmente por hombres blancos que comparten opiniones políticas concretas, incluyendo opiniones sobre política identitaria (Criado Perez, 2019, 23). La tecnocracia tampoco es neutral o apolítica. La competencia técnica no puede zanjar por sí sola debates éticos y políticos. Y las compañías tecnológicas tienen sus propias políticas. Por ejemplo, de acuerdo con Murray (2019, 110), las compañías de Silicon Valley como Google son políticamente de izquierdas (de forma más precisa: libertarias de izquierdas) y esperan lo mismo de sus empleados (incluso si lo que predicán no siempre se practica, por ejemplo cuando se trata de la diversidad de sus propias plantillas). Y las compañías tecnológicas como Amazon o Uber también monitorizan el desempeño de sus empleados mediante IA y algoritmos: Amazon usa IA para despedir automáticamente a los trabajadores con baja productividad (Tangerman, 2019) y los algoritmos de Uber hacen un *ranking* de sus conductores que dicta cuánto ganarán e, incluso, si son o no despedidos (Bernal, 2020). Estas prácticas, en cuanto son explotadoras, contrastan con su discurso político, y son ellas mismas cualquier cosa menos ética o políticamente neutrales. Finalmente, como ya señalé en el capítulo anterior, más allá del entorno inmediato de las compañías tecnológicas y de su mercado, los servicios de IA dependen del trabajo humano del hemisferio sur. Este no cosecha las justas recompensas por su labor.

Dada esta carencia de neutralidad por parte de la IA, de los datos y de las personas y organizaciones que tratan con la tecnología, estas operaciones, prácticas, interpretaciones y percepciones tecnológicas necesitan ser evaluadas. Sin embargo, aún queda abierta la cuestión del *cómo*, esto es, sobre qué base puede llevarse a cabo esta evaluación normativa. Por lo tanto, es de vital importancia que discutamos las normas y los conceptos relevantes. En este capítulo, esto quiere decir que discutimos lo que entendemos por sesgo y discriminación, (in)justicia, (des)igualdad, y así sucesivamente, y por qué exactamente son asuntos problemáticos. He introducido otra parte de un marco conceptual *filosófico-político* normativo para evaluar a la IA que se basa en nociones como la igualdad y la justicia. Esto puede orientar las discusiones sobre política de la IA también en aquellos lugares donde se desarrolla y usa la tecnología.

Lo anterior hace surgir adicionalmente la cuestión de *quién* tiene que evaluar y enfrentarse con los problemas. Los que desarrollan la tecnología tienen aquí un papel importante, ya que son también responsables de los efectos de sus tecnologías. Y ya han adoptado en parte este papel: como empleados estimulados para hacerlo así por las compañías y organizaciones para las que trabajan, las cuales defienden, al menos de boquilla, el desarrollo de una IA responsable; como empresarios interesados en algo más que en hacer dinero; o como ciudadanos que están motivados para cambiar las cosas, por ejemplo como *hackers*. Como muestra Webb (2020), en respuesta a la revolución digital y la vigilancia de masas, la concentración del poder y el autoritarismo que este permite, está teniendo lugar una lucha para perturbar y cambiar las cosas. Hackear es, entonces, parte de un movimiento social y de un nuevo tipo de activismo que hace las cosas por su cuenta. Como ciudadanos, los *hackers* luchan «para recuperar sus democracias» (4), quizás también para preservar la libertad y alcanzar más justicia e igualdad. Más allá de eso, se podría también aspirar a una educación ciudadana más amplia sobre el impacto político de la IA. Esto plantea a continuación el problema de cómo medir ese impacto que, a menudo, es difícil de predecir (Djeffal, 2019, 271). Además, el significado y el impacto político mismos pueden ser discutidos. Se necesita desarrollar más y nuevas herramientas para explorar significados y consecuencias

sociales y políticas, tanto futuras como potenciales, de la IA y para crear las condiciones para una discusión de alta calidad sobre estas posibilidades. Esto sería útil tanto para los que desarrollan la IA como para los ciudadanos.

De forma más general, dadas las enormes y desafiantes discusiones sobre cuestiones políticas fundamentales que surgen en un segundo plano, los trabajadores tecnológicos, las compañías, las organizaciones, los gobiernos, los *hackers*, los profesores y los ciudadanos no deberían quedarse solos para responder al desafío de las políticas de la IA, y debería haber una discusión pública y más amplia sobre el tema. En las sociedades democráticas, nosotros (los ciudadanos) deberíamos decidir sobre la dirección política. Juntos, con la ayuda de expertos, podemos ayudar a aquellos que desarrollan la tecnología para crear las lentes a través de las cuales se pueda analizar el sesgo en la IA y, si fuera necesario, corregirlo dentro y a través de las tecnologías. Las herramientas conceptuales de la filosofía política (en este capítulo, las discusiones sobre la justicia y la igualdad) pueden contribuir a mejorar la calidad de estas discusiones democráticas, a encontrar una dirección normativa y a rediseñar las tecnologías. El siguiente capítulo discute en mayor profundidad la relación entre la IA y la democracia.

CAPÍTULO 4

Democracia: cámaras de eco y totalitarismo de las máquinas

INTRODUCCIÓN: LA IA COMO AMENAZA PARA LA DEMOCRACIA

Junto con los derechos humanos y el Estado de derecho, la democracia es normalmente considerada como uno de los elementos centrales de las constituciones liberales occidentales (Nemitz, 2018, 3) y del pensamiento político liberal. Muchos sistemas políticos de todo el mundo apelan al ideal de la democracia. ¿Podría la IA fortalecerla o, más bien, la debilita? ¿Qué impacto tiene la IA sobre la democracia, dadas sus omnipresentes consecuencias sociales y políticas?

Hoy en día muchos críticos advierten de que la IA es una amenaza para la democracia. En lugar de ayudar a establecer «una nueva era ateniense de la democracia», como dijo el vicepresidente Al Gore sobre la red informática en un discurso a mediados de los años 90 (Saco, 2002, xiii), se teme que las tecnologías de IA conduzcan a un mundo menos democrático, si no distópico, al marcar el inicio de una época en la que internet y el uso de la IA creen un nuevo tipo de *ágora* político. Los críticos cuestionan la idea de que la tecnología sea políticamente neutral, o dudan de que las tecnologías digitales como internet y la IA conduzcan necesariamente hacia el progreso. Por ejemplo, en *La era del capitalismo de la vigilancia* (2019), Zuboff argumenta que el capitalismo de la vigilancia, que usa técnicas de modificación de comportamiento de masas, es una amenaza no solo para la autonomía individual, sino también para la democracia, ya que acaba con la soberanía del pueblo. Para argumentar por qué esto es antidemocrático se refiere a *Los derechos del hombre* de Paine, que nos pone en guardia contra la aristocracia, ya que tal cuerpo de personas se considera inmune a la rendición de cuentas (513). Esta vez, sin embargo, la tiranía no proviene de

los aristócratas, sino del capitalismo de la vigilancia, una forma cruda de capitalismo (518) que expropia la experiencia humana e impone un nuevo tipo de control: una concentración de conocimiento implica una concentración de poder. Webb está de acuerdo en *Los nueve gigantes* (2019), no tenemos el control desde que las grandes corporaciones configuran nuestro futuro. Esto no es democrático. O, como lo expresa Diamond (2019): lo que es bueno para las grandes tecnológicas «no es necesariamente bueno para la democracia, o incluso para nuestra salud individual mental y física» (21). Además, no solo los ciudadanos, sino también los Estados se vuelven cada vez más dependientes de las grandes corporaciones y de lo que estas saben de los ciudadanos (Couldry y Mejias, 2019, 13). Sin embargo, Zuboff (2019) aún cree que es posible una reforma a través de la democracia. Inspirada por Hannah Arendt, cree que es posible un nuevo comienzo, que podemos «reclamar el futuro digital como el hogar de la humanidad» (525). Por el contrario, Hatari defiende en *Homo Deus* (2015) que en el futuro la democracia podría degradarse y desaparecer por completo, ya que no puede lidiar con los datos: «A medida que aumenta tanto el volumen como la velocidad de los datos, podrían convertirse en obsoletas instituciones venerables como las elecciones, los partidos y los parlamentos» (373). El cambio tecnológico es rápido y la política ya no es capaz de seguir el ritmo. Ya está ocurriendo hoy en día que las elecciones cruciales que afectan a nuestras vidas como, por ejemplo, las elecciones relacionadas con internet, no se han realizado a través de procesos democráticos. Incluso los dictadores están abrumados por el ritmo del cambio tecnológico.

Este capítulo investiga adicionalmente el impacto de la IA en la democracia revisando la teoría filosófico-política sobre ella y sus condiciones, incluyendo la relación entre el conocimiento/la pericia y la democracia, e indagando sobre los orígenes del totalitarismo. Primero, bosqueja la tensión entre, por una parte, las concepciones platónicas y tecnocráticas de la política que subrayan el conocimiento, la educación y la pericia de los gobernantes y, por otro lado, los ideales de democracia participativa y deliberativa al estilo de Dewey y Habermas, que, a su vez, han sido criticados por Mouffe y Rancière, quienes ofrecen un ideal

diferente, radical y agonístico de la democracia y la política. Se pregunta sobre cómo puede amenazar o apoyar la IA estos diferentes ideales y concepciones de la democracia. Por ejemplo, la IA podría usarse para posibilitar una democracia directa, pero también podría apoyar una tecnocracia autoritaria: un Gobierno de expertos o (como sugiere Harari) de la propia IA. Y, si la democracia necesita que los ciudadanos tengan información sobre las opiniones de los demás y que participen en una deliberación orientada hacia el consenso, entonces la tecnología digital podría facilitarlo, pero también hay algunos fenómenos que conducen a la fragmentación y polarización de la esfera pública y, así, amenazan este ideal de democracia. Finalmente, la IA se puede utilizar por aquellos que buscan destruir lo político en sí mismo: como una herramienta para los reyes-filósofos platónicos o para la democracia habermasiana, la IA se podría usar para impulsar una comprensión racionalista o tecnológicamente solucionista de la política, que ignore su dimensión inherentemente agonística y corra el riesgo de excluir otros puntos de vista.

Segundo, este capítulo se pregunta si la IA podría contribuir a crear las condiciones para el totalitarismo. Basándonos en la obra de Arendt (2017) sobre los orígenes del totalitarismo, considera si la IA apoya a una sociedad en la que la soledad y la falta de confianza crean un terreno fértil para las tendencias totalitarias. Y, si se usa la IA como una herramienta para la gestión de las personas por parte de las grandes corporaciones y la burocracia gubernamental por medio de los datos, esto significa entonces tratar a las personas como cosas y puede (a través de las, a menudo, imprevistas consecuencias de la gente que tan solo hacen su trabajo) conducir a lo que Arendt (2006) denominó «la banalidad del mal». Este no es solo un problema del pasado histórico, sino que también constituye un peligro presente cuando nuestros datos se encuentran en las manos de personas en empresas tecnológicas y administraciones gubernamentales que pueden hacer lo que les digan que hagan sus gestores o políticos y dejar de *pensar*.

LA IA COMO AMENAZA PARA LA DEMOCRACIA, EL CONOCIMIENTO, LA DELIBERACIÓN Y LA POLÍTICA MISMA

Comenzando con Platón: democracia, conocimiento y experiencia

Para saber si y cómo puede amenazar la IA a la democracia debemos saber primero qué es la democracia. Examinemos diferentes opiniones sobre ella y sus condiciones, incluyendo la relación entre el conocimiento/pericia y la democracia.

La democracia es una respuesta a una antigua pregunta: ¿quién debería gobernar? La teoría de la democracia a menudo contesta a Platón, quien desechó la democracia y argumentó que para gobernar se necesita conocimiento, en concreto conocimiento sobre el bien y la justicia. En *La República*, asocia democracia con ignorancia y argumenta que conduce a la tiranía. Usando la analogía de la navegación, propone que un buen líder debería ser un experto, ya que, como capitán, necesita ser capaz de controlar el barco del Estado. Otra analogía que utiliza Platón es la del médico: si estás enfermo, querrás tener consejo experto. Gobernar el Estado es una cuestión de oficio y pericia. Por lo tanto, deberían gobernar los filósofos, ya que están enamorados del conocimiento y de la búsqueda de la realidad y la verdad. Por «filosofía» Platón no quería referirse a filosofía académica: en *La República*, deja claro que sus guardianes también recibirían formación musical, matemática, militar y física (Wolff, 2016, 68). Otorgar poder a las personas en general (al pueblo) implicaría, por el contrario, dejar que gobernasen la ignorancia, la histeria, el placer y la cobardía. Sin el liderazgo apropiado, el conflicto político y la ignorancia darían como resultado la necesidad de un líder fuerte, un tirano.

Esta visión cambia en la época moderna cuando surgen las nuevas concepciones de la naturaleza humana y de la política de acuerdo con las cuales la mayoría, y no solo unos pocos, son capaces de autogobernarse. Por ejemplo, el pensamiento de Rousseau también busca alejarse de la tiranía, pero para él la solución al problema platónico de la autoridad no es el gobierno de los pocos educados, sino la educación de todos los ciudadanos: el autogobierno es posible y deseable, suponiendo que todos

los ciudadanos reciban educación moral (Rousseau, 1997). Así, Rousseau sentó las bases de otra rama en la teoría política, que cree en la democracia, amplía el autogobierno a todos los ciudadanos y añade algunas condiciones para el autogobierno. Pero ¿qué condiciones debería adoptar este autogobierno y qué forma debería tener? Una cosa es decir que la democracia no debería ser una cuestión de gobierno por reyes-filósofos y otra especificar qué conocimiento, si es que lo hay, es necesario para la democracia, y definir la forma precisa que esta debería tener. Por ejemplo, ¿debería involucrar la deliberación y la participación? Y ¿cómo puede y, de hecho, se relaciona la IA con esta forma de democracia?

Comencemos con la cuestión relacionada con el conocimiento y la democracia. Una discusión que se produce en la teoría política y que es aquí pertinente es el debate tecnocracia/democracia, que comenzó con Platón, si no antes, y que adquirió una nueva relevancia con el surgimiento de las burocracias modernas. Durante las últimas décadas ha habido demandas a favor de que se tomen decisiones orientadas por los datos, de que se practique una gobernanza inteligente y se desarrollen políticas científicas basadas en la evidencia, pero que se encuentran en tensión con las peticiones a favor de un gobierno participativo y de formas más radicales de democracia. Las divergencias entre estos diferentes argumentos y lenguajes (Gilley, 2016) reflejan las opiniones opuestas sobre el papel del conocimiento y la pericia en la política.

A menudo se percibe la IA como algo que está firmemente del lado de la tecnocracia. Ofrece nuevas posibilidades de generar conocimiento sobre la realidad social; se podría decir que para construir la realidad social. La ciencia de la estadística se ha usado ya durante mucho tiempo en la gobernanza moderna, pero con el aprendizaje automático las posibilidades del análisis predictivo se amplían. Couldry y Mejias (2019) hablan de una nueva epistemología social posibilitada por la IA. La IA crea nuevo poder para la dirección tecnocrática de la sociedad que contrasta con los ideales democráticos. La IA parece ser un dominio para expertos, más allá de la comprensión de la mayoría de la gente. Por ejemplo, Pasquale (2019, 1920) ha argumentado que, para una justa distribución de la pericia y del poder, se necesitan crear incentivos para los individuos para comprender la IA y su

cadena de suministros. Si esto no ocurriera, podríamos volvernos totalmente dependientes de la IA y de los burócratas que la utilizan para controlarnos. Transhumanistas como Harari piensan incluso que la IA nos gobernará en el futuro. Pero, dejando de lado a la ciencia ficción, se podría decir que ya estamos gobernados por las grandes corporaciones que usan la IA para manipularnos, todo ello totalmente fuera del control democrático, como ha señalado Zuboff. Quién debería gobernar es meramente una cuestión teórica cuando ya nos encontramos gobernados por Google, Amazon y otros grandes actores corporativos. En este sentido, la IA posee algo esencialmente antidemocrático. Además, ¿es suficiente para la toma de decisiones el tipo de conocimiento que ofrece la IA? Se podría argüir que existe aquí una brecha que requiere el juicio humano y la deliberación democrática. El tipo de inteligencia que exhibe la IA se contrasta a menudo con la inteligencia social humana, que parece necesaria para el discurso político y para dotar de sentido social a las democracias. Pero ¿podría también la IA apoyar formas más democráticas de gobierno y gobernanza y, si así fuera, cómo? Para llevar a cabo esta discusión, necesitamos continuar investigando el problema central de qué es la democracia y qué necesita en términos de conocimiento.

Más allá del gobierno de la mayoría y de la representación

Muchas personas entienden por democracia el gobierno de la mayoría, y piensan en ello como una forma representativa de la democracia. Sin embargo, ambas visiones de la democracia pueden cuestionarse y, de hecho, han sido puestas en cuestión. Para comenzar, no está claro que *la democracia como el gobierno de la mayoría* sea algo bueno necesariamente. Como señala Dworkin (2011): «¿por qué el hecho de que más gente esté numéricamente a favor de una forma de proceder sobre otra es signo de que la política con más apoyo es o más justa o mejor?». Considérese, por ejemplo, que, bajo la influencia de un líder demagógico, una mayoría decide abolir la democracia e instalar un gobierno autoritario: algo sobre lo cual ya nos había advertido Platón. Como mínimo, la democracia como gobierno de la mayoría no parece una condición

suficiente para la democracia: quizás es necesaria, pero se requiere algo más. Algunos añadirían la necesidad de la libertad o de la igualdad, valores que están a menudo consolidados en las constituciones democrático-liberales. Otros añadirían algunos elementos platónicos más, como el requisito de que las decisiones políticas sean buenas (ya que el gobierno de la mayoría no garantiza un buen resultado), lo que, por supuesto, hace surgir la pregunta de qué son una decisión y un resultado buenos; o el requisito de que los gobernantes necesiten de ciertas capacidades y pericia. ¿Deberían ser moralmente buenos, como argumentaban Platón y Rousseau? En cualquier caso, se podría discutir que, incluso en una democracia, son necesarias algunas cualidades de liderazgo. Bell (2016) afirma que el mérito político viene determinado a través de tres atributos: la capacidad intelectual del líder político, sus destrezas sociales (incluyendo inteligencia emocional. Ver Chou, Moffitt y Bryant, 2020) y su virtud. Lo último se encuentra en línea con Platón. Contra la opinión platónica, Estlund (2008) ha argumentado que la concepción platónica de la autoridad política como pericia confunde a los expertos con los jefes. Lo denomina la falacia experto/jefe: algunas personas tienen más pericia que otras, pero «simplemente no se sigue de ello que tengan autoridad sobre nosotros o que deban tenerla» (3). Pero, si separamos la cuestión de la autoridad política de la referente a la pericia, ¿qué significa que tal conocimiento y las cualidades personales no deberían desempeñar ningún papel en absoluto en una democracia? ¿Es absolutamente inevitable la dirección burocrática en el contexto de las naciones Estado tal como las conocemos? ¿Y es también permisible o, incluso, deseable bajo algunas condiciones el uso de la IA para influir en el comportamiento, o conduciría inevitablemente al autoritarismo? Un rechazo a la tecnocracia deja abierta la cuestión de cuál es exactamente el papel que deberían tener los expertos, la pericia y la tecnología en una democracia.

Además, la *democracia representativa* tal como la conocemos también ha sido puesta en cuestión: algunos piensan que solamente la democracia directa es democracia real. Sin embargo, esto parece complicado de poder llevar a cabo en el contexto de (grandes) Estados nación. La antigua versión de la democracia tuvo lugar en una ciudad-Estado, y Rousseau tenía en

mente la ciudad-Estado que conocía (Ginebra). ¿Sería este un nivel mejor para la gobernanza o es posible una democracia directa, incluso en un Estado nación moderno? ¿Podría ayudar la IA en este sentido y, si así es, cómo exactamente?

Existen alternativas a la democracia representativa y al gobierno de la mayoría que, al mismo tiempo, evitan las opiniones platónicas sobre la autoridad política: ideales de la democracia *participativa* y *deliberativa* orientados hacia el logro de un consenso. A veces, estas concepciones de la democracia están formuladas como una respuesta al populismo, que es acusado de menospreciar normas valiosas del debate público «como decir la verdad, tomar en cuenta las razones de otras personas y respetar la evidencia» (Swift, 2019, 96). Esto supone que se necesitan añadir condiciones adicionales a la democracia como gobierno de la mayoría y representación: la idea es que la democracia necesita más de los ciudadanos que simplemente votar de vez en cuando o tener el gobierno de la mayoría. Un significado de democracia, por ejemplo, es la democracia como justificación y deliberación públicas: el debate libre y razonado entre iguales se intenta que alcance el consenso entre los ciudadanos (Christiano y Bajaj, 2021). Una concepción así de la democracia se puede encontrar, por ejemplo, en Habermas, Rawls, Cohen y O'Neill, quienes creen que hay una conexión entre la democracia y el uso de la razón y la deliberación pública. Por ejemplo, en línea con Habermas, Goodin ha defendido en *Reflective Democracy* (2003) una forma de deliberación democrática en la que la gente se imagina a sí misma en la posición de otras personas. En lugar de limitar la democracia al acto «externo» de votar por parte de las personas, él propone centrarse en los actos y procesos «internos» que deberían subyacer a ese acto, en particular, sus juicios reflexivos y sopesados y la determinación conjunta de lo que debería hacerse colectivamente (1). Con «reflexivo» quiere decir, entre otras cosas, que la gente es «más empática con las situaciones difíciles de los demás», también con aquellos distantes y con intereses diferentes, al imaginarse a sí mismos en el lugar de los otros (7). Los hechos y creencias son también importantes, no simplemente lo son los valores (en conflicto) (16).

Así pues, podemos distinguir entre un ideal formal de democracia «fino» o procesual, que otorga a la gente una voz a través del voto, y los ideales «gruesos», que incluyen algunas condiciones como la deliberación, el conocimiento/pericia e imaginación, que hacen de la democracia algo más rico que la «mera agregación de los votos de las personas» (Goodin, 2003, 17). Sin embargo, desde una perspectiva ilustrada, para evitar una elite o versión platónico-autoritaria de la democracia y apoyar formas participativas de la democracia, es importante implicar y educar a todos los ciudadanos; pero incrementar la participación plantea un desafío. Por ejemplo, otorgar a la gente un voto directo no incrementa necesariamente su participación en la política (Tolbert, McNeal y Smith, 2003), y existen muchas formas de actividades políticas disponibles para el público, incluyendo, por ejemplo, la participación política *online* a través de las redes sociales. En cualquier caso, la democracia participativa se toma en serio el propio potencial de las personas para tomar decisiones políticas, en lugar de delegar este punto en el rey-filósofo o en un grupo de elite. Los ideales participativos de democracia se oponen al pesimismo platónico sobre el gobierno de las «masas» e, influenciados por Rousseau y otros pensadores ilustrados, tienen más confianza en los ciudadanos ordinarios y en sus capacidades para la deliberación y la participación política.

¿Qué implica esto para la IA? Este tipo de democracia descarta formas no participativas de gobierno como, por ejemplo, un gobierno *exclusivo* y tecnocrático de expertos y de IA, y con confianza *ciega* y exclusiva en los algoritmos y en las recomendaciones de la IA. Sin embargo, deja abierta la posibilidad de que los expertos y el conocimiento obtenido con la ayuda de la IA puedan, de alguna manera, estar involucrados en el proceso democrático, siempre y cuando los ciudadanos mismos tengan la última palabra y puedan basarse en su propio juicio y discusión. Sin embargo, no está clara la forma precisa que pudiera tener cualquier potencial implicación de la IA. En la práctica, la IA y la ciencia de datos ya se encuentran implicadas en la toma de decisiones democrática pero, ya que la mayoría de los tipos existentes de democracia no son altamente deliberativos ni participativos, es difícil decir cómo podría funcionar esta combinación. Y

continúa existiendo una tensión entre el juicio humano y el cálculo y la predicción de la máquina.

Consideremos más de cerca la teoría de la democracia deliberativa y participativa, pero también a sus críticos radicales.

Democracia deliberativa y participativa contra democracia agonística y radical

En las democracias *deliberativas* los ciudadanos aplican la inteligencia práctica en la deliberación pública entre iguales en la que no solo se preocupan por sus propios objetivos e intereses, sino que también responden ante los de otros (Estlund, 2008). La democracia trata, en este caso, sobre el razonamiento y discusión públicos, y en crear las condiciones para esta discusión (Christiano, 2003). Habermas ve esta discusión como un proceso de comunicación política racional: comunicación guiada por la razón. Su consideración se basa en la conocida «situación de habla ideal», en la que el curso de la deliberación viene determinado solamente por el poder de la razón, libre de influencias coercitivas no racionales y motivada por los deseos de consenso. Más adelante, las presuposiciones de argumentación formaron la base de su ética del discurso (Habermas, 1990). De acuerdo con Estlund (2008), este enfoque introduce valores que van más allá de la democracia misma, ya que trasciende una comprensión procedimental de la democracia. Pero los oponentes de la democracia deliberativa pueden responder que la democracia no se puede reducir a un procedimiento de votación y que necesita implicar el uso público de la razón. La argumentación y la deliberación son, entonces, no algo «adicional» que se añada al ideal de democracia (es decir, otro valor o principio), sino una parte esencial del concepto. Además, tomando en consideración a Goodin y (otros) desarrollos posteriores del pensamiento de Habermas, se podría añadir la adopción de una perspectiva y la solidaridad. Se podría también reconectar la «inteligencia práctica» con el término antiguo *frónesis*: los ciudadanos necesitarían desarrollar entonces la sabiduría práctica, incluyendo (como podemos añadir con Arendt) su imaginación política.

Existen también otras teorías de la democracia que van más allá de ella misma como votación o representación y subrayan la importancia de la comunicación. Por ejemplo, el ideal de Dewey de una democracia participativa necesita de la participación activa de los ciudadanos y pone el énfasis, de nuevo, en el requisito del conocimiento: las personas necesitan educación para poder participar en la política. En *Democracia y educación* (2001), defiende un ideal de democracia que no es simplemente una forma de gobierno, sino un tipo específico de sociedad: «una democracia es más que una forma de gobierno, es principalmente un modo de vivir asociado, de experiencia comunitaria conjunta» (91). Dicho de forma más general, la clave de la política es construir lo social. Sin embargo, para Dewey, la proximidad o trabajar hacia un fin común no es suficiente para que una relación se convierta en realmente social. Reflexiona sobre la analogía de la máquina: las partes de una máquina funcionan juntas para alcanzar un resultado común, pero esto no constituye un grupo social o una comunidad; es todavía demasiado mecánico. Para que exista un ser social y una comunidad reales, la comunicación es esencial. Si queremos trabajar hacia un fin común y alcanzar consenso, necesitamos la comunicación: «cada uno deberá saber en lo que está ocupado el otro y tendrá que tener alguna forma de mantener al otro informado de su propio propósito y progreso» (9). Además, la participación necesita de educación, «que otorga a los individuos un interés personal en las relaciones sociales y su control, y los hábitos mentales que aseguran los cambios sociales sin introducir desorden» (104).

Dewey admite que este ideal de democracia suena platónico, pero rechaza la clase de autoritarismo de este último. Dice que el ideal de Platón estaba «comprometido con hacer de la clase la unidad social en lugar del individuo» (104), y que no dejaba lugar para el cambio. También rechaza el nacionalismo del siglo XIX. Dewey tiene una concepción de la democracia más inclusiva. Se supone que es una forma de vida y una experiencia comunicativa para todos los individuos, no solo para una clase particular. Cree que los individuos pueden ser educados para desarrollar modelos de comportamiento aceptables. Así, para Dewey, la democracia trata de los individuos como seres comunicativos interconectados. Se forja por medio

de la interacción y la comunicación. Requiere de trabajo: mientras que nacemos como seres sociales en el sentido de que estamos asociados unos con otros, la comunidad y la democracia necesitan crearse. Y se crean por todos los ciudadanos, no solamente por un número limitado de representantes.

Sin embargo, la idea de que la democracia debería ser deliberativa y buscar el consenso ha recibido críticas por pensadores radicales como Young, Mouffe y Rancière. En su libro *Inclusion and Democracy* (2000), Young critica especialmente la concepción habermasiana: para ella, la política no trata solamente de argumentos o expresiones imparciales, y la democracia debería ser más *inclusiva* y comunicativa en el sentido de que también debería incorporar nuevas voces y otros estilos y formas de hablar. La deliberación basada en el mejor argumento ignora los estilos de habla y presentación que están subordinados a las formas en las que se expresan las personas educadas. Las normas de diálogo pueden excluir. Las personas pueden también expresarse de una forma diferente, por ejemplo mediante el relato de historias en público (Young, 2000; véase también Martínez-Bascuñán, 2016). Young (2000) defiende una concepción de la inclusión política que trasciende la votación y subraya la comunicación: «La legitimidad normativa de una decisión democrática depende del grado con el que los afectados por ella han sido incluidos en los procesos de toma de decisión» (6). Esto requiere más que derechos de voto: Young defiende que necesitamos considerar diferentes modos de comunicación, representación y organización, y que no deberíamos restringir nuestra concepción de la comunicación política a la argumentación. En las democracias deliberativas se prefieren estilos de expresión particulares, y esto excluye otros estilos y personas específicas. Mientras que ha habido un esfuerzo por interpretar a Habermas (y a Kant) de una forma que haga hueco a las emociones y la retórica (Thorseth, 2008), Young reconoce en consecuencia el papel de las emociones y otros estilos de hacer política.

Mouffe responde a Habermas y a otros ideales de democracia deliberativos y orientados al consenso subrayando la confrontación política y la diferencia: la diferencia permanecerá, y debería permanecer, y no hay esperanza de desembarazarnos del conflicto. Contra los ideales platónicos y

racionalistas de la política y de la democracia, piensa que no existe una solución política duradera para todos los conflictos. El conflicto es, más bien, un signo de que las democracias están vivas. Esto es lo que ella llama la dimensión *agonística* de la política. Entiende la política como «la dimensión antagónica que es inherente a todas las sociedades humanas» (Mouffe, 2016; ver también Mouffe, 1993; 2000; 2005): el conflicto nunca se puede eliminar, independientemente de las prácticas e instituciones políticas específicas. Esto también implica que la exclusión es inevitable. No es posible un consenso racional sin exclusión, un «nosotros» sin un «ellos». El conflicto en las sociedades democráticas no debería ser erradicado. Las identidades políticas están creadas de una forma que delimitan un «nosotros» frente a un «ellos» (ver también el capítulo 6). Además, como Young, Mouffe reconoce el papel de las emociones junto con la razón: las «pasiones» desempeñan también un papel. Sin embargo, el conflicto no implica guerra: los otros no deben ser vistos como enemigos, sino como adversarios, oponentes. Esta comprensión de la democracia reconoce la realidad de la lucha en la vida social (Mouffe, 2016). No existen formas racionales u objetivas de organizar la sociedad: tales soluciones son, también, el resultado de relaciones de poder. El consenso racional es una fantasía. En su lugar, Mouffe propone un pluralismo agonista: un sistema basado sobre el desacuerdo constructivo. Como señala Farkas (2020), Mouffe está influenciada por Wittgenstein, quien pensaba que cualquier acuerdo en torno a opiniones debe basarse sobre el acuerdo en formas de vida. Una fusión de voces no es posible como un producto de la razón, pero sí como una forma de vida común (Mouffe, 2000, 70). En contra de Habermas y de los enfoques deliberativos y racionalistas en general, Mouffe defiende fomentar «una pluralidad de formas de ser un ciudadano democrático» (73). Si fracasamos a la hora de establecer este pluralismo agonista, la alternativa es el autoritarismo, en el que se supone que los líderes tomarán decisiones objetivas y verdaderas.

Rancière también argumenta contra los ideales de la democracia, bien platónicos, bien orientados al consenso, y rechaza tanto la gestión de la democracia por especialistas como su vertiente representativa. Influenciado por una versión particular del socialismo, argumenta que rechazar la

democracia directa es tener una actitud de condescendencia hacia la clase de personas que carecen de educación. En su lugar, propone escuchar lo que tengan que decir los trabajadores. Su versión de la política y de la democracia favorece la acción política basada en el desacuerdo y el disenso. En *El desacuerdo* (1999) y en *Disenso* (2010) argumenta que las manifestaciones de igualdad ocurren cuando las desigualdades incrustadas en el orden actual se interrumpen y reconfiguran. Esto es necesario. Rancière cuestiona si representación y democracia van de la mano. Cree que nuestras instituciones son representativas, pero no democráticas: son oligárquicas. Pero «no hay una buena razón por la que algunos hombres deban gobernar a otros» (Rancière, 2010, 53). El poder de las clases gobernantes debería ser puesto en cuestión. A la democracia no se la debería culpar de la inestabilidad del sistema representativo. Critica la distinción entre la masa ignorante y la elite lúcida y razonable. Ni siquiera una crisis justifica el gobierno de expertos y especialistas. En una entrevista (Confavreux y Rancière, 2020) dice: «nuestros gobiernos han estado operando desde hace tiempo bajo la coartada de una crisis latente que evita que los asuntos mundiales puedan ser confiados a sus habitantes ordinarios y requiere que sigan estando al cargo de especialistas en gestiones de crisis». En cambio, cree que la gente ordinaria es perfectamente capaz de adquirir conocimiento. En *El maestro ignorante* (1991), sugiere que «todos los hombres tienen una inteligencia igual» (18), que los pobres y los privados del derecho al voto pueden ser autodidactas, y que no deberíamos depender de expertos para nuestra emancipación intelectual.

Influenciados por Mouffe y Rancière, Farkas y Schou (2020) han intervenido en el debate sobre los bulos y la posverdad argumentando en contra de hacer equivalentes la idea de democracia con «las ideas de la razón, la racionalidad y la verdad de una forma *a priori*» (5). Cuestionan la opinión de que lo que está amenazando a la democracia es simplemente la falsedad. Contra Habermas, y en la línea de Mouffe y Rancière, consideran a la democracia en constante evolución y que continúa siendo el objeto de lucha política y social. También cuestionan la democracia representativa: la democracia no trata solamente sobre votar, y la razón no puede salvar la democracia. Argumentan también en contra del solucionismo basado en la

verdad, una única fórmula para la comunidad política, el ideal de un consenso racional y la fundamentación del consenso en la verdad y en la razón. Más bien, la democracia ha producido siempre diferentes verdades (en lugar de *la* Verdad con uve mayúscula) y tiene múltiples causas o justificaciones. Si se trata de la verdad, la democracia trata sobre la diferencia, la pluralidad y la multiplicidad:

Lo que continúa siendo propio de una democracia que funciona correctamente no es tanto su capacidad para navegar basada en la razón y en la verdad, sino su capacidad para incluir y dar voz a proyectos y grupos políticos diferentes. La democracia trata sobre diferentes visiones de cómo podría organizarse la sociedad. Trata sobre el afecto, las emociones y los sentimientos (Farkas y Schou, 2020, 7).

Farkas (2020) advierte de que los «bulos» se pueden convertir en un arma retórica para atacar a los oponentes y pregunta: ¿quién marcará la línea entre lo que es un bulo y lo que es real? ¿Quiénes se han de constituir como autoridades? Y, si el significado de la política, de la verdad y de la posverdad tiene carácter *performativo* y es un constructo, uno debería preguntarse: ¿quién lleva a cabo (*performs*) qué discursos de la verdad y por qué? En línea con Derrida, Farkas y Schou (2018) creen que la clausura del significado está basada en la exclusión y que los discursos son siempre el resultado de «fijaciones particulares de significado resultantes de las luchas políticas que han reprimido vías alternativas a lo largo del tiempo» (301). Este enfoque no quiere decir que la pericia ya no tenga cabida en una democracia, sino más bien que tales tensiones continúan siendo una dinámica en las democracias liberales, y la democracia tiene que equilibrar esas fuerzas. Además, las nuevas tecnologías como la IA pueden ayudar a realizar este ideal de democracia. Farkas y Schou (2020, 9) creen que emparejar las tecnologías digitales con formas más participativas e inclusivas de democracia es la única forma de progresar.

Volveré sobre la cuestión de las emociones en el capítulo siguiente. De momento, está claro que existe una tensión entre aquellos (desde Platón hasta Habermas) que apelan a la pericia, la verdad, la razón y el consenso, por un lado, y aquellos que consideran la democracia como lucha o participación directa, por otro: el gobierno de las personas en lugar del gobierno de la razón, como lo expresan Farkas y Schou (2020, 7). Ambos

ideales deliberativo-participativo y agonista van en contra de las advertencias de que la gente ordinaria pueda favorecer el autoritarismo o ser apática (ver, por ejemplo, Dahl, 1956 o Sartori, 1987; como dice Sartori: los votantes raramente actúan, más bien «reaccionan» [123]) y, contra la opinión, defendida por ejemplo por Schumpeter, de que la

gente ordinaria simplemente no es *competente* para comprender las cuestiones que subyacen tras las decisiones políticas, y que, en consecuencia, están contentos cediendo estas decisiones a personas que consideran mejor cualificadas para lidiar con ellas (Miller, 2003, 40; énfasis de Miller).

Pero ¿cómo puede la gente adquirir la competencia? ¿Es suficiente la autoeducación, como cree Rancière, o necesitamos una educación general como propone Dewey? El ideal de democracia de Dewey es no representativo. Pero, incluso en un sistema representativo, se podría defender que la gente necesita cierta educación para ser capaz de elegir mejor a sus representantes. Además, la educación contrarresta el peligro del autoritarismo. La cuestión crucial no sería otorgar más votos a los que tengan mayor educación, como propuso Mill y que está en línea con la visión platónica, sino en educar a todo el mundo.

¿Pueden ayudar la ciencia y la tecnología a superar la ignorancia con una educación así? Depende en gran medida de cómo se use la ciencia y la tecnología y de qué tipo de conocimiento se ofrezca. Se necesitan hechos, pero puede que no consigan cambiar la opinión de las personas, y, como veremos más abajo, la información, por sí sola, no es suficiente. Además, existe el peligro de reemplazar la política por la ciencia, la tecnología y una buena gestión. El uso político de la IA se moviliza o apoya a menudo a aquellos que creen que las luchas y el complejo desorden de la política pueden reducirse a una toma racional de decisiones y a encontrar resultados que sean objetivamente mejores que otros. Existe la tentación platónica de convertir la política en una cuestión sobre la verdad y la pericia filosóficas. En «Dos conceptos de libertad» (ver capítulo 2), Berlin ya nos advertía en contra de la tendencia a convertir todos los problemas políticos en tecnológicos: si todo el mundo está de acuerdo en los fines, entonces la única cuestión pendiente concierne a los medios: pueden estar «establecidos por expertos o por máquinas», convirtiendo así a los Gobiernos en lo que

Saint-Simon denominó «la administración de las cosas» (Berlin, 1997, 191). Además, el conocimiento científico puede muy bien ser necesario para el juicio político, pero, sin duda, no es suficiente. Magnani (2013) ha defendido que el conocimiento que necesitamos en la moralidad no debería ser solamente científico, «sino también humano y social, integrado en tradiciones “culturales” reformuladas», y necesitamos una moralidad «capaz de lidiar con la nueva condición global de la vida humana», teniendo en cuenta que, a menudo, las consecuencias son espacial o temporalmente distantes (68). Este último punto necesita imaginación, y lo mismo se podría decir sobre el conocimiento político. Además, se podría defender que hacer política más inclusiva y menos tecnocrática, y basarse en una inteligencia compartida y colectiva, como proponía Dewey, es una forma de lidiar con estos problemas, ya que incorpora conocimiento social y situado: el conocimiento integrado y conformado por contextos históricos y culturales concretos, que la ciencia no proporciona (¿fácilmente?). Esto podría ayudar a la imaginación política.

Pero, si la democracia necesita ser más deliberativa y participativa, ¿cuál es entonces el papel de la IA? ¿Puede ayudar al tipo de deliberación, comunicación, participación e imaginación concebida por los teóricos que hemos tratado aquí? ¿O es una mera herramienta para manipular a los votantes y, de forma más general, convertir a la gente en cosas (datos)? ¿Puede ayudar al razonamiento público o es una amenaza a ese ideal, dado que podría no ser transparente que haga recomendaciones y cómo haga, y que (en la forma de, al menos, aprendizaje automático) la forma en que trabaja no tiene nada que ver con razonar o juzgar? Y ¿pueden las personas (incluyendo los políticos) tener suficiente conocimiento y destreza como para lidiar con la información ofrecida por la IA y la ciencia de datos, o nos encontramos en manos de una elite tecnocrática que sabe lo que es bueno para nosotros? ¿Qué pasa si la IA ofrece, por emplear la frase de Scanlon, un régimen de gobierno que «nadie podría rechazar razonablemente» (Scanlon, 1998, 153)? ¿Crea principalmente la IA un entorno para los modos de gobierno platónicos o es posible una democracia *à la* Habermas o Dewey? ¿Se usará potencialmente la IA, que está del lado de la razón y de la objetividad, contra las personas que están consideradas como demasiado

emocionales? ¿Podría también desempeñar un papel en una democracia agonista y radical, incluyendo el mantenimiento de la diferencia y el fomento de la emancipación de todos? Una forma de discutir en mayor detalle estas cuestiones es señalar al problema de las burbujas de información y las cámaras de eco, a las cuales nos dirigimos ahora.

BURBUJAS DE INFORMACIÓN, CÁMARAS DE ECO Y POPULISMO

Basándose en estos ideales de democracia deliberativa y agonista, se podría proponer que las tecnologías inteligentes como la IA deberían contribuir a facilitar una participación más amplia e inclusiva a través de las redes sociales. Pero, entonces, deberíamos considerar los desafíos y limitaciones de estas tecnologías en términos de conocimiento.

Algunas de estas cuestiones sobre internet y las redes sociales ya se conocían en los estudios sobre medios de comunicación. Por ejemplo, Sunstein (2001) analizó problemas relacionados con la personalización, fragmentación y polarización, y Pariser (2011) defendió que la personalización crea una burbuja-filtro que limita nuestro horizonte. Pero ahora la combinación de las redes sociales y la IA exacerba estas tendencias. La cuestión se ha formulado en términos de burbujas de información y cámaras de eco (véase, por ejemplo, Niyazov, 2019): los algoritmos de personalización alimentan a la gente con información con la que probablemente interactúan, y el resultado es la segregación en burbujas donde sus propias creencias están reforzadas y donde no se exponen a opiniones opuestas. Esta situación hace que la imaginación política que concebía Goodin (2003) se vuelva más complicada: parece que obstaculiza las formas empáticas de política en lugar de promoverlas. Hay también una mayor polarización política, que torna imposible tanto el consenso como la acción colectiva. Hay también un riesgo de fragmentación y ruptura de la sociedad. Las redes sociales están consideradas como especialmente problemáticas: mientras que las publicaciones impresas, la televisión y la radio tienen cámaras de eco propias, aún «ejercen algún tipo de control editorial» (Diamond, 2019, 22). Esto está ausente en las cámaras de eco de las redes sociales, lo que conduce a la polarización y al lenguaje del odio.

Esto se vuelve problemático si lo que se quiere es disponer de un debate racional orientado hacia el consenso, como en el caso de Habermas. Pero incluso una forma de democracia agonista es complicada cuando las personas no están expuestas a opiniones opuestas. El-Bermawy (2016) ha defendido que la aldea global ha sido sustituida por «islas digitales de aislamiento que se van alejando a la deriva cada vez más». Hay más segregación. En Facebook consumimos principalmente contenido político similar a nuestros puntos de vista. En este sentido, desarrollamos una visión en túnel. Nguyen (2020) distingue entre burbujas epistémicas, que omiten voces relevantes (a menudo de forma no intencionada), y cámaras de eco, que son una estructura de la que se han excluido activamente otras voces relevantes y en las que la gente desconfía de todas las fuentes exteriores. Los motores de búsqueda, por ejemplo, pueden atrapar a los usuarios en filtros burbuja y cámaras de eco mediante la forma en que funcionan, amenazando la diversidad y, por tanto, la democracia (Granka, 2010). Ciertamente hay un riesgo, a pesar de que el trabajo empírico detecta que las redes sociales también exponen a los usuarios a perspectivas opuestas y que solo un pequeño grupo de usuarios busca deliberadamente entornos de opinión al estilo de cámaras de eco (Puschmann, 2018). Las redes sociales también permiten a la gente publicar opiniones que serían censuradas en la mayoría de los principales canales de noticias. En este sentido, existe, al menos, una oportunidad para la diversidad de opiniones.

Sin embargo, de acuerdo con los ideales deliberativos de la democracia, esta no consiste simplemente en un intercambio de opiniones y va más allá de preocupaciones temporales. Por ejemplo, el ideal deliberativo de democracia tiene que ver con el uso público de la razón e incluye deliberar cómo vivir juntos durante un largo periodo de tiempo y adquirir compromisos. Benhabib (entrevistado en Wahl-Jorgensen, 2008) argumenta en contra de las opiniones que reducen la democracia al «intercambio de opinión sin trabas» a través de internet y olvidan los «compromisos de actuación» a largo plazo (965), como dar parte del salario a la comunidad. La IA también puede amenazar a la racionalidad comunicativa y a la esfera pública concebida por Habermas, Arendt y otros. De acuerdo con Benhabib, es un desafío «conceptualizar la interacción entre estas redes de

comunicación, información y formación de opiniones, por un lado, y la articulación pública en términos de articulación en términos de decisiones, por otra» (964).

Formulado en términos de conocimiento, se podría defender que las cámaras de eco amenazan la fundamentación epistémica de la democracia (al menos, de acuerdo con las consideraciones deliberativas, participativas y republicanas de la democracia). Kinkead y Douglas (2020) plantean el problema de la forma siguiente: los pensadores políticos desde Rousseau y Mill hasta Habermas, Goodin y Estlund creen en el poder epistémico, la virtud y la justificación de la democracia, ya que el debate público libre nos permite rastrear la verdad y compartir y discutir una diversidad de opiniones (121). Sin embargo, las redes sociales, combinadas con análisis de *big data*, cambian la naturaleza de la comunicación política: en lugar de exponer las propias opiniones en la discusión pública y el escrutinio mediante la radiodifusión, ahora es posible enviar mensajes altamente orientados a muchas personas en todo el mundo, para «llevar a cabo una retransmisión selectiva (*narrowcasting*) con alcance global» (129). Este aspecto tiene consecuencias epistémicas:

Un riesgo para la virtud epistémica de la democracia es que las redes sociales cerradas se apropien de la esfera pública y la conviertan en privada. Una vez que es privada y compartida solamente entre individuos similares, el discurso político pierde algo de su robustez epistémica, ya que las ideas ya no se desafían por parte de perspectivas diferentes (127).

Además, en el discurso privado es más sencillo usar la manipulación sin que los participantes se den cuenta (127-128). Considérese de nuevo la opinión de Mill de que las mejores ideas y la verdad surgirán en un mercado abierto de ideas: esta apertura se encuentra amenazada por las cámaras de eco, los filtros-burbuja y la retransmisión selectiva.

De forma más general, se podría defender que esta privatización de la esfera pública es altamente problemática desde un punto de vista democrático, ya que la democracia necesita una esfera *pública* y la política trata sobre asuntos públicos. Los fenómenos como las cámaras de eco amenazan estos puntos. Pero ¿qué es la esfera pública y qué es «lo público», a la luz de las tecnologías digitales actuales? La distinción entre lo público y lo privado parece obsoleta cuando la gente comparte sus

pensamientos y sentimientos más privados en las redes sociales y la tecnología y la política basadas en la identidad también ejercen presión sobre esta distinción. No obstante, teniendo en cuenta las razones ofrecidas por las teorías deliberativas de la democracia, es bueno retener y defender una noción de «lo público». De forma adicional, se podría señalar la necesidad de soluciones colectivas a problemas colectivos. Couldry, Livingstone y Markham (2007) escriben que como la ciudadanía no es simplemente una elección de estilo de vida, las cuestiones públicas y la política

implican más que una simple «pertenencia social» o ciertas expresiones de identidad [ver también de nuevo los argumentos en contra de la política identitaria]. En la distinción entre lo público y lo privado algo crucial sigue en juego, independientemente de los pensadores posmodernos que celebran su fracaso (6).

Argumentan que la democracia necesita tener una noción de algunas cuestiones que son públicas y que necesitan una definición común y una solución colectiva. Pero la tecnología puede no hacer fácil la distinción entre las preocupaciones públicas y las privadas. Y la crítica política puede, a menudo, quedar excluida de la esfera pública llamada «política» y tener lugar en otras esferas como el arte y, ciertamente, las redes sociales, que pueden ser, o no, públicas, o públicas solo en algunos aspectos.

Una cuestión relacionada con lo anterior es el populismo, que puede estar conectado con la IA de varias formas. Por ejemplo, los políticos populistas usan la IA para analizar datos sobre las preferencias de los votantes. Pero, a pesar de que es bueno en una democracia que los políticos sepan lo que necesitan los ciudadanos, este uso de la IA «podría convertirse en llamamientos de masas demagógicos, en lugar de los procesos de deliberación razonada previstos, por ejemplo, por los padres fundadores de los Estados Unidos» (Niyazov, 2019). Algunos teóricos son más positivos con respecto al populismo. De acuerdo con Laclau (2005), el populismo no es simplemente una realidad óptica, empírica (un tipo particular de política), sino que es sinónimo de política: trata de la política como tal. Él ve la posibilidad de un proyecto político renovado. Pero la mayoría de las teorías aquí consideradas, y de movimientos como el movimiento *Occupy*, se distancian del populismo. En cualquier caso, en las redes sociales vemos

la valorización y glorificación de los comentarios de los que no pertenecen a las elites y un rechazo del conocimiento experto (Moffitt, 2016), que se ha llamado «populismo epistemológico» (Saurette y Gunster, 2011). Moffitt (2016) ve en esto aspectos positivos, por ejemplo el sacar a la luz la corrupción, pero también señala a las cámaras de eco. Las redes sociales pueden contribuir a la propagación del populismo y de una mayor división ideológica y favorecer la viralidad e inmediatez de la política de la discusión y la comprensión. Mientras que las redes sociales no están necesariamente dirigidas por la IA, esta puede desempeñar un papel en este aspecto filtrando y regulando la información y en forma de bots, que influyen en la comunicación política y, potencialmente, también en las preferencias de los votantes. Esto podría preparar la sociedad para un líder autoritario que afirme que encarna la voluntad del pueblo y cuyas obsesiones personales se conviertan en las de la nación. La IA, a través de las redes sociales, puede contribuir al auge del populismo y, eventualmente, al autoritarismo.

Más problemas: manipulación, sustitución, responsabilidad y poder

La IA se puede emplear para manipular a la gente. Ya he mencionado las posibilidades del *nudge*, que influye en las decisiones (ver el capítulo 2). Como otras tecnologías digitales, puede incluso usarse para moldear la experiencia y el pensamiento humanos. Como señala Lanier (2010), hablando en nombre de los tecnólogos:

Nos entretenemos con tu filosofía mediante la manipulación directa de tu experiencia cognitiva, no indirectamente, por medio de una argumentación. Solo se necesita un pequeño grupo de ingenieros para crear tecnología que pueda moldear el futuro entero de la experiencia humana con increíble velocidad (7).

En el contexto de la democracia representativa y de sus procedimientos de voto, la IA y otras tecnologías digitales pueden empujar a los votantes para que apoyen a un político o a un partido concreto a través de anuncios personalizados, y así sucesivamente. Niyazov (2019) argumenta que esto puede conducir a la tiranía de la minoría (los votantes dubitativos). Un caso conocido que implicaba la manipulación de las personas es el de Cambridge

Analytica (ver el capítulo 2), que incluía la recolección de datos privados de los usuarios de Facebook sin su permiso. Se ha afirmado que esos datos fueron usados a continuación con el propósito de influir en procesos políticos, como, por ejemplo, la campaña presidencial de Donald Trump de 2016. Teniendo en mente los criterios de Bell (2016) para el liderazgo político, el uso manipulador de la IA puede también reflejar una tendencia que se aleja de la inteligencia, de las destrezas sociales y de la virtud. Y si la IA llegase a tomar el liderazgo político, es dudoso que la IA pudiera alcanzar nunca las capacidades intelectuales, las destrezas sociales y la virtud necesaria, de acuerdo con Bell.

El reemplazo del liderazgo humano por la IA es, en cualquier caso, peligroso y antidemocrático. El peligro no es solamente que una IA que tome el control pueda destruir la humanidad, sino también que gobierne, o afirme que gobierna, *teniendo en cuenta el mejor interés de la humanidad*. Este es un tema recurrente en la ciencia ficción clásica (véase, por ejemplo, la película *Yo, robot* o las novelas *Polity* de Neal Asher) y un peligro de acuerdo con todos los ideales de la democracia: destruye la democracia liberal, ya que destruye la libertad como autonomía de los ciudadanos y, al final, nos arrebatara la política como tal, como ha defendido Damnjanović (2015). Este escenario solo podría apoyarse en una interpretación de Platón muy peculiar en la que la IA actuaría como un rey-filósofo artificial (quizás, combinado con un argumento utilitarista moderno que busque maximizar la utilidad para la humanidad como un todo o un argumento hobbesiano que busque la supervivencia y la paz de la humanidad). El capitán o piloto (*kybernetes*) del barco del Estado se convertiría en un piloto automático. Estos puntos de vista han sido correctamente criticados por todas las teorías democráticas que hemos considerado aquí. Tampoco está claro si una IA así podría imponer *de facto* su autoridad, independientemente de si es correcto y justo tener una tal autoridad proveniente de una IA.

Desde el punto de vista de una filosofía de la tecnología, tales escenarios y discusiones muestran que la IA no es meramente una tecnología, sino que siempre está preparada para estar conectada con posibilidades morales y políticas. Tales posibilidades ilustran la no instrumentalidad de la IA: la IA no es simplemente una herramienta para la política, sino que cambia la

política misma. Además, incluso si la IA se usa dentro de un marco democrático, existe un problema de responsabilidad y legitimidad cuando las decisiones se toman en función de recomendaciones de la IA, especialmente si no es transparente cómo se tomaron esas decisiones o si se introdujeron o reprodujeron sesgos durante el proceso. Un ejemplo bien conocido es la IA que ayuda a los jueces de los Estados Unidos en cuestiones de condenas penales, libertad condicional y elegibilidad para servicios sociales (ver el caso COMPAS mencionado en el capítulo 3). Para una democracia, la responsabilidad pública es crucial: los funcionarios públicos que toman decisiones sobre los ciudadanos deberían ser responsables, y no está claro cómo esto se puede mantener si a la hora de tomar decisiones dependen en un alto grado de la IA, y si la forma en que esta alcanza esas decisiones no es transparente ni neutral. El último punto es también importante, ya que, como se mencionó al final del capítulo 2, se puede considerar la igualdad como una condición para la democracia. Desde esta perspectiva, es también altamente problemático el que estemos controlados, en su caso, por una industria tecnológica monopolizada que controla los datos y sus flujos y, por tanto, a las personas en última instancia. Nemitz (2018) ha criticado la actual concentración del poder digital. Defiende que los desafíos planteados por la IA no pueden abordarse solamente por la ética de la IA, sino que es necesario que se afronten con leyes aplicables y legítimas que sean el resultado de un proceso democrático. Reclama una cultura de la IA que sea democrática.

El poder de los gigantes tecnológicos también plantea un problema cuando se trata de lidiar con los bulos y la desinformación. Mientras que los bulos (información falsa o engañosa presentada como noticias) se sacan también a colación en las críticas tecnodeterministas y abiertamente pesimistas de las tecnologías de las redes digitales (Farkas y Schou, 2018, 302), estos fenómenos plantean un problema político serio también con relación a la democracia. ¿Qué cuenta como «bulos» y quién lo decide? ¿Deberían las empresas de redes sociales como Twitter y Facebook convertirse en censores, en aquellos que deciden qué contenido se permite y cuál no, en lugar de, por ejemplo, que lo decida un cuerpo elegido democráticamente? Y, si es así, ¿cómo deberían lidiar con este asunto?

Como hemos señalado, las empresas de internet usan la moderación de contenidos: esto es, censura. Dado el tamaño de las plataformas de redes sociales actuales y de las pequeñas ventanas temporales, esta automatización parcial o completa de la moderación de contenidos parece inevitable. Pero esto implica que las reglas del discurso público (¡y las reglas de lo que cuenta como hechos y verdad!) están establecidas por un pequeño grupo de elites de Silicon Valley; esto equivale a un tipo de gobernanza oculta, lejos del foco de las instituciones democráticas. De nuevo, hay una falta de transparencia y de responsabilidad. Tampoco está claro cómo se debe lidiar con cuestiones complejas que conciernen a la política, e incluso existe el riesgo de que estén despolitizadas (Gorwa, Binns y Katzenbach, 2020). Además, los algoritmos pueden dejar que prospere el discurso de odio, pero pueden también restringir la libertad de los usuarios para acceder a los contenidos y expresar sus opiniones. No está claro si los derechos de los usuarios se encuentran suficientemente protegidos. Y la regulación del discurso hecha por algoritmos no refleja necesariamente las transacciones sociales. Las comprobaciones y los equilibrios para asegurar que el poder se ejerce en favor de la sociedad en su conjunto también están ausentes (Elkin-Koren, 2020). Las capacidades de la IA se encuentran limitadas a la hora de lidiar de forma efectiva con esta cuestión (y con todas las demás cuestiones políticas complejas). Además, desde una perspectiva marxista, se podría decir que la libertad de expresión misma se convierte en una mercancía, algo que se transforma en un valor económico en la economía de datos. ¿Qué significan la libertad y la participación política en este contexto? Y ¿quién define los términos de acuerdo con los cuales se lleva a cabo la discusión pública? Un pequeño número de actores poderosos establecen los términos y condiciones. Los denominados «acuerdos» de los usuarios, mediante los cuales estos aceptan los términos definidos por plataformas como Facebook o Twitter, no son constituciones democráticas, sino imposiciones. Esta cuestión subraya de nuevo que la IA no funciona únicamente como una herramienta políticamente neutral en manos de los humanos que participan en un juego político particular, sino que transforma el juego mismo, transforma las condiciones bajo las cuales se lleva a cabo la política. La IA media y

conforma la propia democracia. Niyazov (2019) cree que las sociedades abiertas pueden lidiar con desafíos que afectan a la democracia y a la igualdad, ya que proporcionan vías para el pensamiento crítico. Pero ¿qué pasa si las sociedades abiertas se transforman en otra cosa, en algo mucho más opaco y hostil hacia el pensamiento crítico?

Nótese también que en las democracias necesitan realizarse difíciles transacciones entre distintos valores políticos, por ejemplo entre libertad e igualdad. Como vimos en el capítulo 2, Tocqueville vio una tensión fundamental entre ellos. Estaba muy preocupado por el hecho de que la igualdad debilitase las garantías de la libertad individual y de los derechos de las minorías, y que esto nos condujera potencialmente a la tiranía. Rousseau, por otro lado, pensaba que la democracia y la verdadera libertad necesitaban de igualdad política y moral, que, a su vez, necesitaban de una mínima calidad socioeconómica. Actualmente continúa habiendo debates en teoría política sobre la igualdad, por ejemplo en respuesta a la obra de Piketty. En el capítulo 3 hemos visto que la IA puede afectar a la igualdad de formas diversas. El éxito y la legitimidad de las democracias también depende de que sean capaces de negociar y equilibrar diferentes principios políticos, que pueden estar afectados por la forma en que la IA ya influye en la realización de estos valores de una forma no transparente y no pública. Por ejemplo, si la democracia necesita de igualdad moral, como defendía Rousseau, entonces esto ya es razón suficiente para reducir la desigualdad social basándose en estos motivos, sin tener en cuenta otros valores. Si la IA aumenta la desigualdad porque, por ejemplo, conduzca al desempleo o aumente el sesgo, entonces, de acuerdo con este razonamiento, esto sería antidemocrático, ya que incrementa la desigualdad *moral y política*. Sin embargo, la democracia también necesita defender la libertad incluyendo, por ejemplo, un grado suficiente de libertad negativa. Si una redistribución de los bienes conduce a una menor libertad negativa, entonces se necesita encontrar una negociación aceptable entre libertad e igualdad. Este equilibrio se hace cada vez más complicado una vez que consideramos diferentes nociones de la justicia, que puede que no siempre tengan como objetivo la igualdad como tal, sino que, por ejemplo, primen el tratamiento preferencial (actual o previo) de grupos desfavorecidos. Las justificaciones

de por qué la IA es antidemocrática pueden depender entonces de otros valores como la igualdad y la justicia, que pueden estar en tensión entre sí. Las apelaciones a la democracia, en el contexto de las discusiones sobre la IA y sobre cualquier otro tema, no pueden evitar las complicadas políticas de estas negociaciones y equilibrios.

Para finalizar, lo que aquí vemos no es solamente que la IA puede usarse como una herramienta para socavar directamente la democracia. También posee efectos colaterales imprevistos, como hacer que el ideal de democracia como deliberación sea más difícil de alcanzar, reforzar el populismo, amenazar la responsabilidad pública e incrementar la concentración de poder. Además, las políticas relativas a la IA deben equilibrar valores políticos diferentes que están potencialmente en conflicto.

LA IA Y LOS ORÍGENES DEL TOTALITARISMO: LECCIONES DE ARENDT

IA y totalitarismo

La democracia también puede definirse en contraste con el autoritarismo y el totalitarismo. Este último no solo es autoritario por disponer de un poder central fuerte que viola considerablemente la democracia en cualquier sentido de los subrayados hasta ahora (el voto, la participación ciudadana, el pluralismo y la diversidad, etc.), sino que también interviene profundamente en las esferas pública y privada de sus ciudadanos. Se caracteriza por la represión política, la censura, la vigilancia de las masas, el terrorismo de Estado y la ausencia total de libertad pública. Son ejemplos históricos de regímenes totalitarios la Alemania nazi en tiempos de Hitler, la Unión Soviética en los de Stalin, la China comunista de Mao y la Italia fascista de Mussolini. Actualmente, las tecnologías digitales ofrecen nuevos medios de vigilancia y manipulación que pueden apoyar o conducir al totalitarismo. La IA es una de estas tecnologías. No solo puede ayudar a los gobernantes autoritarios y a sus partidarios a amañar las elecciones, a diseminar desinformación y a controlar y reprimir a la oposición; también puede ayudar a crear un tipo particular de vigilancia y control: la *vigilancia total* y el *control total*. Bloom (2019) advierte de la amenaza del

«totalitarismo 4.0», que conduce a una situación en la que «todos serán completamente analizados y contabilizados. Todas sus acciones serán monitorizadas, conocida toda preferencia y su vida entera calculada y predecible» (vii). Y si la vigilancia mediante la IA conduce al conocimiento sobre todo el mundo, puede afirmarse que la IA nos conoce mejor de lo que lo hacemos nosotros mismos. Esto abre el camino para el paternalismo (capítulo 2) y el autoritarismo. Como señala McCarthy-Jones (2020):

Las sociedades occidentales individualistas están construidas sobre la idea de que nadie conoce nuestros pensamientos, deseos o alegrías mejor que nosotros mismos. [...] La inteligencia artificial (IA) cambiará esta situación. Nos conocerá mejor de lo que nos conocemos nosotros mismos. Un Gobierno armado de una IA podría afirmar conocer todo lo que su pueblo quiere realmente.

McCarthy-Jones compara esta idea con lo que ocurrió en la Unión Soviética de Stalin y en la China de Mao. La IA facilita que dispongamos de una versión digital del Gran Hermano en la que cada ciudadano se encuentra bajo vigilancia constante mediante «telepantallas». Actualmente esto suena bastante familiar, especialmente, pero no de forma exclusiva, en Estados con tendencias autoritarias y totalitarias. Tecnológicamente hay poco que obstaculice la vigilancia de masas por parte del Estado usando IA y ciencia de datos. Considérese el sistema de crédito social de China, basado en huellas digitales dejadas por los individuos. El Estado utiliza capturas de vídeo con cámaras de vigilancia, *software* de reconocimiento facial, reconocimiento de voz y datos privados de empresas tecnológicas como Alibaba y Baidu. Diamond (2019, 23) llama a este caso una forma de «totalitarismo posmoderno».

Sin embargo, la IA no solo se usa para apoyar el totalitarismo de Estado; también permite un tipo de totalitarismo de las grandes corporaciones. Considérese, de nuevo, la afirmación de Zuboff de que vivimos bajo un «capitalismo de la vigilancia» (Zuboff, 2015; 2019): una nueva lógica de acumulación capitalista que monitoriza y modifica el comportamiento humano. En lugar de referirse al Gran Hermano, Zuboff (2015) habla sobre el «Gran Otro»: en lugar de un control centralizado del Estado, nos enfrentamos a «un régimen institucional conectado ubicuamente que graba, modifica y mercantiliza la experiencia cotidiana desde tostadores hasta

cuerpos, desde la comunicación hasta el pensamiento, todo con vistas a establecer nuevas formas de monetización y beneficio» (81). Este hecho puede que ya esté ocurriendo hasta cierto punto a través de las redes sociales, pero en la medida en que el «internet de las cosas» y las tecnologías relacionadas transforman nuestros hogares, nuestros puestos de trabajo y nuestras ciudades en entornos inteligentes, también podemos imaginar fácilmente cómo podrían convertirse cada vez más en lugares en los que todo lo que está ocurriendo se encuentra bajo el ojo vigilante de la tecnología electrónica, incluyendo la IA. Y la IA no solo nos observa, sino que también realiza predicciones acerca de nuestro comportamiento. La IA y la ciencia de datos pueden convertirse así en un instrumento de nuevas formas de totalitarismo en las que la IA nos conozca mejor (y *antes*) que nosotros. Como ha señalado Harari en una entrevista en *WIRED* (Thompson, Harari y Harris, 2018), los sentimientos y las elecciones humanos ya no son lugares sagrados. Los seres humanos pueden ser ahora manipulados completamente: «ahora somos animales hackeables». Esto abre nuevas posibilidades para la tiranía, tanto por los Gobiernos como por las grandes corporaciones. El conocimiento obtenido mediante la IA y tecnologías relacionadas se puede utilizar para manipularnos y controlarnos.

Así pues, la tecnología se arriesga a convertirse (usando una frase de Arendt (2017)) en uno de los *orígenes del totalitarismo*. El cambio de la democracia al totalitarismo no ocurre (simplemente) porque un *Führer* o un presidente asuma y destruya abiertamente la democracia mediante, por ejemplo, una revolución o un golpe de Estado. Más bien, el proceso es mucho menos visible y lento, pero no menos efectivo. Mediante la IA y otras tecnologías electrónicas, el equilibrio del poder se está inclinando poco a poco hacia las manos de unos pocos actores poderosos (tanto en el Gobierno como en las grandes corporaciones) y, ciertamente, alejándose del pueblo, si es que, para empezar, han tenido siquiera algo de poder. En este sentido, la IA es más que una mera herramienta: es un cambio radical. En la medida en que se aplica al campo político también cambia ese campo, por ejemplo cuando contribuye a la creación de una dinámica totalitaria.

Sin embargo, la tecnología no es el único factor que pone en peligro la democracia, y sus efectos no se materializan de forma aislada del entorno

humano y social en el que opera y en el que se utiliza. A la IA por sí sola no se la puede culpar, y no debería culpársela, de los peligros aquí indicados. El entorno humano y social, y las formas en las que estos interactúan con la tecnología, son, al menos, igual de importantes. Para comprender este punto, necesitamos extraer lecciones de aquellos que analizaron casos históricos de autoritarismo y totalitarismo, que se preguntaron cómo las democracias podrían deteriorarse hasta alcanzar su propio opuesto, y aquellos que estaban motivados por la opinión normativa de que lo que ha ocurrido nunca debería volver a ocurrir. En el siglo anterior muchos intelectuales se hicieron precisamente estas preguntas tras la Segunda Guerra Mundial, y Arendt fue una de ellos.

Arendt sobre los orígenes del totalitarismo y de la banalidad del mal

En *Los orígenes del totalitarismo* (2017), publicado originalmente en 1951 y escrito teniendo como trasfondo el totalitarismo de la Alemania nazi y de la Unión Soviética, Arendt no solo describe formas concretas de totalitarismo; también presenta una investigación sobre las condiciones que preparan a la sociedad para su llegada. Defiende que si los movimientos totalitarios, con «su capacidad insuperable para establecer y salvaguardar un mundo ficticio a través de una mentira consistente» (499) y su «desdén por toda la textura de la realidad» (xi), tienen éxito a la hora de establecer una transformación totalitaria de la sociedad, es porque ya existían condiciones asentadas que preparaban a la sociedad moderna para ello. Ella señala la incapacidad de los modernos de vivir en, y dotar de sentido a, el mundo que ellos mismos han creado con un poder incluso mayor (xi). En particular, subraya cómo la soledad («aislamiento y falta de relaciones sociales normales» [415]) convierte a la gente en vulnerable a las formas violentas de nacionalismo, que son usadas a continuación por líderes totalitarios como Hitler. De forma más general, «el terror puede gobernar absolutamente solo sobre hombres que se encuentran aislados entre sí» (623). Considérese de nuevo el razonamiento de Hobbes: solo cuando las vidas de los individuos son «desagradables, brutales y cortas», cuando se encuentran aislados y compiten entre ellos, puede el leviatán instaurar su

gobierno autoritario de la espada. El problema no es el totalitarismo y el autoritarismo como tales, existe una herida más profunda: la ausencia de solidaridad y de acción colectiva y, en última instancia, la destrucción de la esfera política misma. Como escribe Arendt: «el aislamiento es ese *impasse* al que se empuja a los hombres cuando la esfera política de sus vidas, cuando su actuación conjunta en búsqueda de una preocupación común, se destruye» (623). Esto indica que estamos ante un mundo sin confianza, un mundo «en el que nadie es de fiar y en el que no se puede contar con nada» (628).

Actualmente, el fenómeno de los movimientos que se protegen a sí mismos y a sus seguidores «del impacto de la facticidad» (Arendt, 2017, 549) suena bastante familiar tanto cuando consideramos el trumpismo, los bulos y el terrorismo (no gubernamental) como cuando se habla de soledad y de incapacidad de otorgar sentido al mundo. Este no es solo un problema de las personas menos educadas o de los excluidos de la sociedad. Muchos simpatizantes de Trump son de clase media (Rensch, 2019), y ciertamente no todos ellos son, o fueron, personas solitarias en el sentido de estar solos o de no tener amigos, pero pueden ser vistos como *políticamente* solitarios en el sentido de Arendt; carecen de un mundo de solidaridad y confianza. Reconociendo la relevancia de cuestiones como el sesgo, la explotación y el neocolonialismo (ver capítulos 3 y 5), y el rol de la propaganda de ideología populista y de derechas, se podría decir que, hoy día, en los Estados Unidos el aislamiento político y un mundo desprovisto de confianza ayudan a conformar un terreno ideal para el surgimiento del autoritarismo. Si Arendt está en lo cierto, el autoritarismo y el totalitarismo no *crean* tanto un tejido social dañado como hacen crecer un tejido que *ya* está dañado. Visto desde la perspectiva de Arendt, el totalitarismo es, en sentido estricto, no un movimiento político, sino un movimiento que destruye la propia esfera política. No es simplemente antidemocrático en el sentido de autoritario, sino que es también una suerte de «soledad organizada» (Arendt, 2017, 628): la destrucción de la confianza en los demás y la corrosión de la creencia en la verdad y en los hechos. Bajo esta perspectiva, debemos hacernos de nuevo la pregunta relacionada con la tecnología: ¿pueden

contribuir, o contribuyen de forma efectiva, a crear estas condiciones las tecnologías contemporáneas como la IA? Y si es así, ¿cómo?

Ciertamente, la IA puede utilizarse para proteger a la gente de la realidad, o para crear una distorsión de la realidad, por ejemplo cuando crea las burbujas epistémicas anteriormente mencionadas o, directamente, propaga la desinformación. Pero también puede contribuir a las condiciones psicosociales y socioepistémicas subyacentes para el surgimiento del totalitarismo mencionadas por Arendt. Una versión de este argumento podría presentarse centrándonos en la soledad literal (diferiendo en cierto punto de Arendt). De acuerdo con Turkle, las máquinas pueden contribuir a la soledad, ya que solamente nos facilitan la ilusión de la compañía. En *Alone Together* (2011), escribe que los robots «pueden ofrecer la ilusión de la compañía sin las exigencias de la amistad» (1). Estamos conectados, pero nos sentimos «completamente solos» (154). Podríamos dejar ya de correr el riesgo de la amistad con humanos, ya que tememos la dependencia asociada con ella (66). Nos ocultamos detrás de la pantalla, e incluso telefonar a otras personas es visto como algo demasiado directo. Pero, si hacemos esto, perdemos la empatía humana y el cuidado de los unos por los otros. Perdemos la capacidad de responder ante las necesidades de los demás. Perdemos la amistad y el amor reales y nos arriesgamos a tratar a los otros como si fueran objetos, usándolos para nuestra comodidad o diversión (154). La amplitud de estos riesgos y la cuestión de si son creados por las redes sociales es un asunto controvertido. Turkle puede que sea demasiado despectiva frente a las posibilidades sociales positivas que ofrece la tecnología, pero el peligro de estar «solos estando juntos» necesita tomarse en serio. Si Arendt está en lo cierto, esta soledad no es solo un triste estado a nivel personal, sino que conduce a una pérdida generalizada de confianza y solidaridad; es también un problema *político* y peligroso hasta el punto de que ayuda a abonar el terreno para el totalitarismo.

En la medida en que la IA y otras tecnologías digitales contribuyen a estos fenómenos, aquellas son políticamente problemáticas. Considérese, por ejemplo, cómo las redes sociales crean ansiedad y pueden resultar en una tribalización: somos bombardeados constantemente por malas noticias sensacionalistas y solo confiamos en la información de nuestra «tribu»

(Javanbakht, 2020). Las cámaras de eco y las burbujas epistémicas contribuyen aún más a esta tribalización. La ansiedad exagera la soledad y la separación, y la tribalización puede no solo convertirse en polarización política y distribución de la esfera pública, sino también conducir a la violencia. Además, cuando Arendt (2017, 573-574) afirma que los campos de concentración nazi transformaron a las personas humanas en meras cosas de formas científicamente controladas, se podría también considerar cómo las formas contemporáneas de manipulación del comportamiento a través de la IA pueden tener un impacto moral similar en la medida en que conviertan a la gente en lo que Arendt denomina animales «pervertidos», abiertos al hackeo. Considérese de nuevo el capitalismo de la vigilancia y cómo la economía de datos está basada en la explotación y la manipulación de las personas. Y existen más paralelismos entre las atrocidades que ocurrieron en la Segunda Guerra Mundial y, de forma general, los males del totalitarismo. En *Privacidad es poder* (2020), Carissa Véliz compara el escenario de un régimen autoritario contemporáneo que se apropia de nuestros datos personales con el uso de registros por parte de los nazis para asesinar judíos en masa (115). Como ella misma lo describe: «la recopilación de datos puede matar» (114). La IA y la ciencia de datos se pueden usar para tales propósitos y, de forma menos visible, para crear las *condiciones* para tales fenómenos.

Sin embargo, nunca es la tecnología por sí sola la que mata o instala y mantiene un régimen totalitario. Se necesita también a las personas: en particular, gente que obedezca órdenes; gente que no oponga resistencia. Y esto nos conduce hasta otro libro de Arendt: *Eichmann en Jerusalén* (2006), famoso por su subtítulo: «Un estudio sobre la banalidad del mal». Escrito en 1963, Arendt analiza el juicio que tuvo lugar dos años antes de Adolf Eichmann, un nazi que desempeñó un papel importante en el asesinato en masa de judíos durante la Segunda Guerra Mundial. La propia Arendt fue testigo del juicio. El estudio provocó un montón de reacciones. En lugar de ver a Eichmann como a un monstruo, o como alguien que odiase a los judíos, Arendt se tomó en serio su propia opinión de que había seguido órdenes: «Habría tenido mala conciencia solo si no hubiera hecho lo que le habían ordenado que hiciera» (25). Cumplió con su «deber» (135), en el

sentido de obedecer órdenes y seguir la ley de la Alemania nazi, que implicaba seguir las órdenes del *Führer*. No podía haber excepciones (137), era una «virtud» ser obediente (247). Este análisis contribuyó al proyecto de Arendt de investigar los orígenes del totalitarismo: concluyó que la obediencia, la no resistencia, es parte del mal implicado. La Alemania nazi no habría podido cometer sus crímenes y atrocidades sin las muchas personas que, como Eichmann, «simplemente» seguían órdenes y buscaban proseguir su carrera profesional. Esta es la cara banal y ordinaria del mal totalitarista, pero no por ello menos «temible» (252). Para Arendt, esta obediencia era más importante que la vida interior y las motivaciones de los líderes totalitarios (278). Sin embargo, tiene la esperanza de que siempre habrá alguien que se resista: «Bajo condiciones de terror la mayoría de la gente acatará las órdenes, pero otra parte de ella no» (233).

Para comprender la relación entre la IA y el totalitarismo debemos no solo examinar las intenciones y motivaciones de la gente (que pueden ser buenas o malas, como, por ejemplo, intentar manipular a las personas para así alzarse con el poder, o banales, como, por ejemplo, hacer carrera en una compañía de IA como un científico de datos), sino también considerar las consecuencias no intencionadas y cómo *hacer simplemente tu trabajo* puede contribuir a estas consecuencias. Normalmente, el sesgo no es intencionado. Por ejemplo, es bastante probable que un equipo concreto de desarrolladores y científicos de datos no *pretendan* incrementar el sesgo en la sociedad, pero, al realizar su trabajo dentro de una gran empresa u organización gubernamental, puede que hagan eso exactamente. A pesar de que puede haber un pequeño número de personas con malas intenciones (en empresas tecnológicas o en cualquier otro lugar), normalmente no es este el caso. Más bien, haciendo simplemente tu trabajo y obedeciendo a la autoridad se puede contribuir a la creación o proliferación del sesgo. Desde una perspectiva arendtiana, el mal se encuentra en no cuestionar, en no reflexionar, en hacer simplemente lo que se supone que tenemos que hacer. Se encuentra localizado en la banalidad de la gente que cumple con su «deber» en prácticas tecnológicas diarias y en las estructuras jerárquicas relacionadas con ellas. El mal se encuentra en el momento en que uno otorga la conformidad cuando lo que se necesita es una no conformidad

para evitar un resultado malo o dañino o, políticamente hablando, surge en el momento en el que no se ofrece resistencia cuando resistir es hacer lo correcto.

La resistencia no es solo importante bajo el totalitarismo, también lo es en la democracia. En parte, la posibilidad de la desobediencia se encuentra consagrada en el mismo marco legal. Como defiende Hildebrandt (2015): «En una democracia constitucional la desobediencia y la contestación son las marcas distintivas de la ley» (10). En una democracia en la que rige el Estado de derecho, los ciudadanos pueden rebatir normas y sus aplicaciones. Pero con Arendt se podría ir más allá y defender de forma más controvertida que, *independientemente* de lo que diga una ley (o un tribunal judicial), la resistencia puede estar justificada y ser necesaria en función de motivos morales y políticos. En cualquier caso, la cuestión crucial que defiende Arendt consiste en que seguir reglas y órdenes *ciegamente* es muy peligroso y moralmente problemático y que este asunto es también relevante en las democracias.

Un argumento relacionado señala la falta de *reflexión*. Inspirado por *Eichmann en Jerusalén*, McQuillan (2019) argumenta que, ya que la IA ofrece a la gente «rankings de riesgo empíricos, cuya derivación no tienen forma de cuestionar», la tecnología «fomenta la inconsciencia en el sentido descrito por Hannah Arendt: la incapacidad de criticar a las instituciones, la falta de reflexión sobre las consecuencias, el compromiso con la creencia de que se lleva a cabo una orden correcta» (165). El daño del enfoque estadístico promovido por la IA es también que, en la medida en que está basado en datos históricos, obtenemos más de lo mismo, nos mantenemos estancados en lo viejo. En *La condición humana* (1958) Arendt escribe:

Lo nuevo siempre ocurre contra las abrumadoras posibilidades de las leyes estadísticas y su probabilidad, que para todos los propósitos prácticos del día a día equivale a la certeza. Lo nuevo, por lo tanto, aparece siempre bajo la apariencia de un milagro. El hecho de que el hombre sea capaz de actuar significa que lo inesperado puede esperarse de él, que es capaz de realizar lo que es infinitamente improbable (178).

Mientras que la forma de McQuillan de considerar la IA suena demasiado determinista y se arriesga innecesariamente al asumir una oposición estricta entre los humanos y la tecnología (los humanos también

desempeñan un papel), existe un peligro importante de que ciertos usos concretos de la IA (esto es, ciertas combinaciones concretas de IA y humanos) contribuyan a dar forma a las condiciones bajo las cuales puede crecer y florecer el totalitarismo.

La cuestión sobre las condiciones es importante. Para evitar el totalitarismo de la IA (y para mantener la democracia), no es suficiente señalar la responsabilidad de las personas en empresas tecnológicas y en organizaciones gubernamentales y decir que deberían mejorar el diseño de la tecnología, los datos, y así sucesivamente. También es necesario hacerse las siguientes preguntas: ¿qué tipo de entorno social podría crearse que apoye a la gente a la hora de ejercer esta responsabilidad y les facilite cuestionar, criticar o, incluso, oponer resistencia, cuando la resistencia es hacer lo correcto? ¿Qué barreras pueden crearse para obstaculizar los cambios descritos desde la democracia hasta el totalitarismo? Y ¿cómo podemos crear las condiciones bajo las cuales puede florecer la *democracia*?

La respuesta a estas preguntas depende, por supuesto, del ideal de democracia (y de política) que uno contemple. En este capítulo he facilitado una perspectiva general de algunos de estos ideales y de las tensiones entre ellos, pero es necesario llevar a cabo más trabajo para investigar las condiciones que hacen funcionar a la democracia. Aquí, la filosofía y la ciencia (¡y las artes!) pueden colaborar. Por ejemplo, inspirado en el volumen de Miessen y Ritts (2019) sobre políticas espaciales del populismo de derechas, nos podemos preguntar sobre las condiciones espaciales y materiales para la democracia. ¿Qué espacios serían buenos para la deliberación democrática? Y ¿qué tipo de espacios podría producir la IA? ¿Cómo podría ayudar la IA a crear, literal y metafóricamente, buenas arquitecturas para la democracia? ¿Qué tipo de *ágora* y espacio público necesitamos, no solo conceptualmente; sino también de forma concreta, material y espacialmente? Y ¿cuál es la relación entre, por una parte, lo político y lo social y, por otra, los artefactos materiales? Por ejemplo, he preguntado en otra parte (Coeckelbergh, 2009a) qué implicaría hablar de una «política de los artefactos»: cuestioné la definición antropocéntrica de Arendt de la *polis* en *La condición humana*, que acepta una estricta

distinción entre humanos y artefactos, pero subrayé la importancia de los acontecimientos políticos (el lanzamiento del Sputnik en su prólogo) y recuperé la idea de que para un ámbito público necesitamos un mundo común que nos reúna a todos, quizás incluyendo de alguna manera una «comunidad de las cosas» (Arendt, 1958, 52-55). En el capítulo 6 discutiré en más profundidad la idea de incluir a no humanos dentro de la esfera política y el papel de lo híbrido en la política. De forma más general, necesitamos saber más sobre las relaciones precisas entre, por una parte, la política y lo social y, por otra, el conocimiento, el espacio y las tecnologías materiales. Necesitamos reflexionar sobre las condiciones para la democracia y la construcción de la esfera pública a la luz de las nuevas tecnologías y entornos tecnológicos.

Una forma en la que pueden contribuir a este proyecto la filosofía política y la filosofía de la tecnología es conceptualizando el poder (y su mucho menos comprendida) relación con la tecnología. Si queremos comprender mejor lo que estamos haciendo y lo que *podríamos* hacer juntos a la luz de tecnologías como la IA (no simplemente los peligros, sino también las oportunidades), necesitamos entender cómo funciona el poder y cómo se relaciona con el conocimiento y la tecnología. Este es el tema del siguiente capítulo.

CAPÍTULO 5

El poder: la vigilancia y la (auto)disciplina por medio de los datos

INTRODUCCIÓN: EL PODER COMO TEMA DE LA FILOSOFÍA POLÍTICA

Una forma de hablar sobre política es utilizando el concepto de poder. El poder es a menudo considerado en términos negativos o como una representación de cómo son realmente las cosas, en cuanto opuestas a un ideal. Por ejemplo, el poder se ha invocado en respuesta a aquellos que defienden los ideales deliberativos y participativos de la democracia liberal. Considérese de nuevo el ideal de Dewey de una democracia participativa. Los críticos han defendido que este ideal es ingenuo, ya que evita hablar sobre el conflicto y el poder. En particular, es considerado demasiado optimista por lo que respecta a la capacidad de los ciudadanos ordinarios para juzgar y actuar de forma inteligente y sobre las probabilidades de lograr un consenso, ignorando, por tanto, lo que Hildreth (2009) llama «las fuerzas oscuras de la naturaleza humana, incluyendo la sed de poder y la voluntad de manipular las relaciones sociales para el propio beneficio» (781). No mucho después de Dewey, Mills escribió en *La elite del poder* (1956) que la sociedad americana está gobernada por miembros de grandes corporaciones, integrantes del ámbito militar y del Gobierno, que «están al mando de las principales jerarquías y organizaciones de la sociedad moderna» (4) y tienen acceso al poder y la riqueza de las que allí disponen. En lugar de ciudadanos «sometidos a un control responsable por una pluralidad de asociaciones voluntarias que conectan el debate público con su culminación en términos de decisiones», como los defensores de la democracia participativa pueden imaginar que deberían funcionar las cosas, Mills vio un «sistema de irresponsabilidad organizada» (361) dirigida por una elite. La resolución pública de problemas, como la imaginó Dewey, no

funciona a gran escala. La política necesita de la lucha por el poder, y no se puede modular de acuerdo con modelos científicos de resolución de problemas. Dewey ignoró equivocadamente cómo se distribuye el poder en la sociedad y cuán profundamente divididas pueden estar las sociedades. Como hemos visto en el capítulo anterior, esta crítica se encuentra también en línea con Mouffe y Ranci  , quienes por el contrario proponen examinar el poder como disenso y agonismo. Y el marxismo cuestiona la distribuci  n de poder entre clases sociales, subrayando c  mo el capital otorga poder a aquellos que lo poseen. En ambos casos, el poder se encuentra relacionado con el conflicto, que puede usarse de forma productiva bajo condiciones espec  ficas.

Otro ejemplo del poder enfrentado a los ideales, que es directamente relevante para la IA, es el poder contra la libertad como consentimiento. En los Estados Unidos y en Europa, pinchar en el lugar donde se indica que aceptas los t  rminos de servicio de una plataforma de internet en concreto (incluyendo su pol  tica de procesamiento de datos y, por tanto, la forma en la que funciona la IA) tiene como objetivo proteger los derechos de los consumidores, incluyendo su libertad. Sin embargo, como ha defendido Bietti (2020), este dispositivo regulatorio no logra justificar las condiciones de trasfondo injustas y las estructuras de poder bajo las cuales tienen lugar estos actos de consentimiento individual. Si los dise  quilibrios de poder est  n «conformando el entorno en el que se toma la decisi  n de otorgar un consentimiento», entonces el consentimiento es «un constructo vac  o» (315). El poder es tambi  n considerado como un peligro para la verdad (Lukes, 2019) y como algo potencialmente enga  oso. El poder se puede utilizar para la coerci  n, por ejemplo, en el contexto de un Estado totalitario, pero tambi  n puede tomar la forma de varios modos de manipulaci  n. Esto amenaza el razonamiento y el desarrollo de las capacidades cr  ticas. Reflexionar es tambi  n dif  cil cuando te encuentras constantemente en un entorno competitivo en el que no te puedes permitir tomarte tiempo para relajarte (Berardi, 2017). El poder es visto entonces como el enemigo de la reflexi  n misma.

Sin embargo, el poder no es necesariamente malo. Una opini  n m  s influyente y, probablemente, m  s compleja del poder es la ofrecida por

Foucault. Inspirado en Nietzsche, Foucault (1981, 93-4) conceptualizó a la sociedad en términos de poder, en particular en términos de relaciones de fuerzas, pero su punto de vista difiere considerablemente del marxismo. En lugar de analizar el poder desde arriba hacia abajo, en términos de soberanía centralizada y el poder de los gobernantes y las elites, Foucault propone un enfoque de abajo arriba, comenzando por los pequeños mecanismos y operaciones de poder que conforman a los sujetos, producen tipos particulares de cuerpos e impregnan la sociedad al completo. Analiza estos micromecanismos de poder en las prisiones y los hospitales. En lugar de relacionar el poder con la cabeza del Leviatán, el soberano central y autoritario en el pensamiento de Hobbes, Foucault (1980) se centra en la pluralidad y en el cuerpo del poder: sobre «la miríada de cuerpos que están constituidos como sujetos periféricos como resultado de los efectos del poder» (98), sobre los «mecanismos infinitesimales» del poder (99). El poder se ejerce dentro del cuerpo social «en lugar de por encima de él» (39), «circula» (98) a través del cuerpo social (119). Además, Foucault está interesado en cómo el poder «alcanza la fibra misma de los individuos, toca sus cuerpos y se introduce en sus acciones y actitudes, sus discursos, sus procesos de aprendizaje y sus vidas cotidianas» (39). Los individuos no son solo los puntos sobre los que opera el poder; más bien, simultáneamente ejercen el poder y se someten a él: son los «vehículos del poder, no sus puntos de aplicación» (98). El individuo es el efecto del poder.

¿Qué implican estas diferentes opiniones sobre el poder para la política de la IA? ¿Es usada la IA por aquellos que manipulan las relaciones sociales para su propio beneficio y nos engañan? Y ¿cómo interactúa con los micromecanismos de poder descritos por Foucault? ¿Qué tipo de individuos, sujetos y cuerpos se crean mediante la IA? En este capítulo, formulo estas preguntas y aplico teoría política y social del poder a la IA. Primero, usaré un marco conceptual general sobre el poder y la tecnología como el desarrollado por Sattarov para distinguir entre diferentes formas en las que la IA puede impactar sobre el poder. A continuación, me serviré de tres teorías del poder para elaborar algunas relaciones entre la IA y el poder: el marxismo y la teoría crítica, Foucault y Butler, y un enfoque orientado hacia la *performatividad* como el propuesto en mi propio trabajo. Esto nos

conducirá a una conclusión sobre lo que yo llamo «poder artificial» (el título inicial de este libro).

EL PODER Y LA IA: HACIA UN MARCO CONCEPTUAL GENERAL

La relación entre la política y la tecnología ya es un tema bien conocido en la filosofía contemporánea de la tecnología. Considérese la obra de Winner (1980), que muestra que las tecnologías pueden tener consecuencias políticas imprevistas, y la teoría crítica de la tecnología de Feenberg (1991), que no solo está inspirada por Marx y la teoría crítica (en particular, Marcuse), sino que también está empíricamente orientada. Sin embargo, mientras que hay por doquier mucho interés en el *poder*, por ejemplo en los estudios culturales, en los estudios de género, en el poshumanismo, y así sucesivamente, ha habido poca atención filosófica sistemática y un número reducido de visiones panorámicas sobre el tema en la filosofía de la tecnología. En la ética de la computación existen trabajos sobre el poder de los algoritmos (Lash, 2007; Yeung, 2016), pero se echa en falta que durante décadas no haya habido un marco sistemático para reflexionar sobre el poder y la tecnología. Una excepción es *Power and Technology* de Sattarov (2019), que distingue entre diferentes concepciones del poder y las aplica a la tecnología. Mientras que su contribución está dirigida principalmente hacia la ética de la tecnología más que hacia la filosofía política de la tecnología, es, sin embargo, muy útil para el propósito de analizar las relaciones entre la IA y el poder.

Sattarov distingue entre cuatro concepciones del poder. La primera, que denomina *episódica*, trata sobre las relaciones en las que un agente ejerce el poder sobre otro, por ejemplo mediante la seducción, la coerción o la manipulación. La segunda define el poder como una *disposición*: como una capacidad, habilidad o potencial. La tercera, una concepción sistemática, entiende el poder como una propiedad de las *instituciones* sociales y políticas. La cuarta ve el poder como *constituyente o productor* de los agentes sociales mismos (Sattarov, 2019, 100). Las dos últimas son, así, más estructurales, mientras que las primeras tratan sobre los agentes y sus acciones (13).

Siguiendo a Sattarov, podemos establecer una correspondencia entre estas diferentes concepciones del poder y las relaciones entre el poder y la tecnología. En primer lugar, la tecnología puede (ayudar a) seducir, coaccionar, forzar o manipular a la gente, y se puede utilizar para ejercer la autoridad. Se podría también decir que este tipo de poder es un poder que se ha delegado a la tecnología o, empleando un concepto a menudo utilizado en la posfenomenología de la tecnología, que la tecnología «media». Por ejemplo, la publicidad *online* puede seducir a los usuarios para que visiten una página web, los resaltes en la calzada pueden obligar a los conductores a reducir la velocidad, y la tecnología puede también manipular. La tecnología puede actuar como «*nudge*»: puede cambiar la arquitectura de nuestras elecciones de manera que sea más probable que las personas se comporten de ciertas formas sin que se den cuenta de ello (ver también el capítulo 2). En segundo lugar, la tecnología puede dar poder a las personas en el sentido de que incrementa sus capacidades y potencial para la acción: puede *empoderar*. Esto es también cierto para la humanidad en general, como ha defendido Jonas (1984): la tecnología ha otorgado a la humanidad un enorme poder. Considérese el concepto del antropoceno: la humanidad en su conjunto se ha convertido en un tipo de fuerza geológica (Crutzen, 2006). Ha adquirido una hiperagencia que ha transformado la superficie entera de la Tierra (ver también el capítulo siguiente). En tercer lugar, cuando se trata del poder sistémico, podemos ver cómo la tecnología puede apoyar sistemas e ideologías particulares. Por ejemplo, desde una perspectiva marxista, la tecnología apoya el avance del capitalismo. El poder, en este caso, no es solo lo que hacen los individuos, sino que más bien está incrustado en un sistema político, económico o social al que contribuye la tecnología. Por ejemplo, los medios de comunicación de masas conforman la opinión pública. Esto es también cierto en el caso de las redes sociales, que pueden apoyar un sistema político-económico concreto (por ejemplo, el capitalismo). Finalmente, si el poder no es simplemente algo que se posea o se ejerza por los individuos, o se aplique a ellos, sino que es también algo constitutivo de los sujetos, yoes e identidades, como defendía Foucault, entonces la tecnología puede usarse para constituir tales sujetos, yoes e identidades. A menudo no existe

ninguna intención de parte de los que desarrollan la tecnología y de los usuarios de hacerlo, pero puede ocurrir igualmente. Por ejemplo, las redes sociales pueden conformar tu identidad, incluso si no eres consciente de ello.

¿Qué significa todo esto para la reflexión sobre el poder y la IA? Primero, la IA puede seducir, coaccionar o manipular, por ejemplo a través de las redes sociales y de sistemas de recomendación. Como los algoritmos en general (Sattarov, 2019, 100), la IA puede diseñarse con el objeto de cambiar la actitud y los comportamientos de los usuarios. Sin usar la coacción, puede funcionar como una «tecnología persuasiva» (Fogg, 2003) mediante la seducción y la manipulación de la gente. Los sistemas que recomiendan música como Spotify, o páginas web como Amazon, buscan dirigir el comportamiento de los oyentes o de los compradores mediante empujoncitos (*nudges*) que cambien el entorno de decisión (ver también los capítulos previos), por ejemplo sugiriendo que otras personas con un gusto similar sobre libros ha comprado el libro *x* y el libro *y*. Y el orden de las publicaciones de Facebook se decide por un algoritmo que puede, por ejemplo, influir en los sentimientos de los usuarios a través de procesos de «contagio» (Papacharissi, 2015). Los individuos se concentran en grupos basados en intereses y comportamientos similares, que pueden reproducir estereotipos sociales y reafirmar antiguas estructuras de poder (Bartoletti, 2020). La gente está también manipulada mediante precios dinámicos y otras técnicas «personalizadas», explotando así las vulnerabilidades de la toma individual de decisiones, incluyendo sesgos bien conocidos (Susser, Roessler y Nissenbaum, 2019, 12). Como en todas las formas de manipulación, la gente está influenciada para actuar de una cierta forma sin que sean conscientes de esta influencia. Como hemos visto, esa influencia encubierta en la toma de decisiones de los individuos amenaza la libertad entendida como autonomía individual. En la medida en que esto suceda dejamos de tener el control de nuestras elecciones, y ni siquiera entendemos los mecanismos subyacentes de cómo ocurre esto. Si bien la concepción moderna de la autonomía de acuerdo con la cual somos o deberíamos ser individuos atomistas y racionalistas no es adecuada y ha sido criticada dentro y fuera de la filosofía occidental dominante (véase, por ejemplo, las

discusiones sobre autonomía relacional en Christman, 2004 y Westlund, 2009), sin embargo, incluso como seres sociales y relacionales queremos algún tipo de control sobre nuestras decisiones y nuestras vidas, y no queremos ser manipulados. En términos de poder, las formas mencionadas de seducción y manipulación de la IA desplazan el equilibrio de poder (más incluso) hacia aquellos que recogen, poseen y monetizan nuestros datos. Además, ciertos grupos particulares de la sociedad (por ejemplo, grupos racistas) pueden intentar obtener poder manipulando a gente en las redes sociales.

Segundo, la IA puede empoderar mediante el incremento de las capacidades individuales de las personas. Considérese, por ejemplo, el procesamiento de los lenguajes naturales que ayuda a traducir y, así, abre nuevas posibilidades para los individuos (a la vez que crea problemas, por ejemplo de descalificación y amenazas a la privacidad). Pero la IA también incrementa el potencial para ejercer el poder sobre los demás, tanto humanos como no humanos, y al final incrementa el poder de la humanidad sobre el entorno natural y la Tierra. Considérese, por ejemplo, los motores de búsqueda y las redes sociales, que pueden empoderar a los individuos que previamente no disponían de acceso a la información en esta cantidad y con este ancho de banda y que podrían no haber tenido voz en los medios de comunicación clásicos. Al mismo tiempo, se les otorga mucho poder a esos motores de búsqueda y a las compañías que los ofrecen: dan forma a los flujos de información y, por tanto, desempeñan el papel conocido como «portero». Además, las compañías y sus algoritmos usan la personalización: «filtran la información individuo por individuo», lo que introduce sesgos humanos y técnicos (Bozdag, 2013, 1). Este papel de portero y estos sesgos conllevan implicaciones para la democracia y la diversidad (Granka, 2010). Como ya hemos visto en el capítulo 3, la IA, en términos de poder, sirve aquí a los intereses de algunas personas más que a los de otras. La IA también se puede usar al nivel del Estado para facilitar la vigilancia y su uso autoritario. Ofrece poder a los Gobiernos y a sus agencias de inteligencia, en el sentido de que los equipa con nuevos instrumentos y capacidades de vigilancia que pueden conducir a aumentar la opresión e, incluso, al totalitarismo. A veces, los Estados y las empresas privadas se

alían para incrementar esas capacidades, como sucede en China y en los Estados Unidos. El sector empresarial tecnológico sabe mucho sobre las vidas de los ciudadanos (Couldry y Mejias, 2019, 13). Incluso en las democracias liberales se están instalando sistemas de reconocimiento facial, el uso de policía predictiva y el empleo de herramientas de IA en sus fronteras. Existe el riesgo de lo que Sætra (2020, 4) denomina una nueva forma de «gobernanza algorítmica» que organice «la acción humana en general». Además, la IA también empodera a la humanidad en su conjunto, lo que puede tener consecuencias para los no humanos, tales como los animales y los entornos naturales. Si, en el contexto del antropoceno, la IA continúa incrementando la capacidad humana para intervenir en la naturaleza y transformarla, entonces estaría apoyando un desplazamiento, ya en curso, en términos de poder: desde los no humanos hacia los humanos. Téngase en cuenta que la IA ayuda a extraer recursos naturales de la tierra y el consumo energético de la tecnología de IA (ver capítulo 6) que, a su vez, también necesita del uso de recursos naturales. Que la IA facilite poder a los humanos puede causar su empoderamiento a nivel individual, pero puede tener graves consecuencias para la naturaleza no humana dado el incremento de los poderes baconianos de la humanidad para minar y transformar la Tierra: el conocimiento científico y las tecnologías son usadas para controlar la naturaleza. En el siguiente capítulo diré más cosas sobre estos aspectos no humanos y terrestres de la política y del poder de la IA.

Tercero, la IA puede apoyar versiones neoliberales del capitalismo, del autoritarismo y de otros sistemas e ideologías. Los sistemas de *software* y *hardware* relacionados con la IA «forman parte de una realidad institucional más amplia de carácter social, económico y político» (Sattarov, 2019, 102), incluyendo aquí sistemas e ideologías socioeconómicas. Estos sistemas más amplios afectan al desarrollo de la tecnología creando, por ejemplo, un contexto de inversión en IA, pero la tecnología también puede ayudar a mantener esos sistemas. Por ejemplo, Dyer-Witheford, Kjosen y Steinhoff (2019) afirman que la IA es un instrumento del capital y que, por lo tanto, implica la explotación y la concentración de poder en manos de los propietarios de los gigantes tecnológicos que, a su vez, están concentrados

en países y regiones concretas, como los Estados Unidos (Nemitz, 2018). Así pues, la IA no es simplemente tecnológica, sino que también crea o mantiene un orden social particular, en este caso, el capitalismo y el neoliberalismo. Considérese también de nuevo la afirmación de Zuboff sobre el capitalismo de la vigilancia: la cuestión crucial no es solo que una tecnología particular sea problemática. La IA y el *big data* ayudan a crear, mantener y expandir todo un sistema socioeconómico en el que el capital es acumulado (por algunos) a través de tecnologías que recolectan y venden los datos (de la mayoría), explotando así la naturaleza humana y alcanzando la esfera de la intimidad. Incluso nuestras emociones son monitorizadas y monetizadas (McStay, 2018). La propia tecnología de IA puede usarse para apoyar regímenes totalitarios o para mantener sistemas políticos opresivos y sus correspondientes narrativas e imágenes (por ejemplo, una utopía racista), a pesar de que, en principio, la IA también podría ofrecer oportunidades para apoyar a la democracia: todo esto depende en gran medida de cómo pensemos la democracia (ver capítulo 4) y, ciertamente, también la política.

La mayoría de los investigadores en IA y en su política apoyan una forma de programación democrática y justa. Algunos creen que se necesitan más límites y más regulación. Los efectos opresivos no son siempre intencionados (y ni siquiera lo son normalmente) de forma intencional, a pesar de que algunas veces la IA se use para promover políticas racistas o nacionalistas. Sin embargo, como hemos visto en los capítulos 3 y 4, existen también efectos problemáticos imprevistos. La IA puede apoyar culturas y sistemas políticos racistas y neocoloniales introduciendo sesgos contra individuos y grupos particulares, o ayudar a crear las condiciones para el autoritarismo y el totalitarismo. Considérese de nuevo el argumento de Noble (2018) de que los algoritmos (de búsqueda) y los sistemas de clasificación pueden «reforzar las relaciones sociales opresivas» (1). Un ejemplo de tal «opresión algorítmica» (4) fue el caso de Google Photos, que etiquetaba a los afroamericanos como «simios» y «animales», un problema que Google realmente no pudo solucionar (Simonite, 2018). Sin embargo, no está siempre tan claro como en este caso si un uso o resultado particular de la IA está sesgado o es injusto, y la cuestión también depende de las

concepciones que cada uno tenga de la justicia y de la igualdad (ver capítulo 3). En cualquier caso, las decisiones, los pensamientos, las acciones y las emociones pueden también estar controladas a propósito para apoyar un sistema político particular. En el caso del totalitarismo, la IA puede dar soporte al alcance ilimitado del sistema con el objeto de penetrar en las mentes y los corazones de las personas.

Cuarto, la IA puede desempeñar un papel en la constitución del yo y en la formación del sujeto, incluso si no somos conscientes de ello. La cuestión aquí es que la IA no solo nos manipula y se inmiscuye profundamente en el nivel personal de un modo que puede ayudar a deducir los pensamientos y sentimientos de las personas, haciendo inferencias sobre sus estados internos tomando como base tanto los comportamientos observables como las expresiones faciales y las preferencias musicales, que son usadas a menudo para la predicción y la monetización por parte del capitalismo de la vigilancia, sino también que la IA contribuye a conformar cómo nos entendemos y nos concebimos a nosotros mismos. Si bien Rouvroy (2013) exige que la «gobernanza algorítmica» evite «cualquier encuentro con sujetos humanos reflexivos» (144), sin embargo, no deja espacio para los juicios humanos y las evaluaciones explícitas de nuestras creencias y de nosotros mismos, y conduce a la explotación de las relaciones entre individuos (Stiegler, 2019), esto no implica que no haya un efecto en nuestro yo (autoconocimiento). ¿Qué tipo de percepción y de conocimiento del yo ayuda a crear la IA? Por ejemplo, ¿comenzamos a entendernos a nosotros mismos como productores y recolectores de datos para su venta? ¿Nos cuantificamos a nosotros y a nuestras vidas a medida que nos rastreamos a nosotros mismos y somos rastreados por otros? ¿Pensamos que tenemos «dobles en términos de datos» (Lyon, 2014), modelos digitales de nosotros mismos, incluso si una IA no almacena un modelo digital del usuario (Matzner, 2019)? ¿Adquirimos y comunicamos un sentido del yo en términos de redes sociales (Papacharissi, 2011)? ¿Qué tipo de identidad y de subjetividad es la que permite la IA?

Realizar tales preguntas va más allá de una comprensión instrumentalista de las relaciones humano-tecnología. La subjetividad del yo y la humana no son algo externo a las tecnologías de la información como la IA. Al

contrario, «la tecnología digital afecta de alguna manera a la propia subjetividad humana» (Matzner, 2019, 109). Las tecnologías de IA han influido en la forma en la que percibimos y actuamos en el mundo, conduciéndonos a nuevas formas de subjetividad (118). Y existen formas diferentes de subjetividad conectadas con la IA. Por ejemplo, basándonos en los tipos de sujetos que somos y el tipo de comunidades a las que pertenecemos, reaccionaremos de forma diferente a sistemas de seguridad concretos basados en IA. El que alguien no sea reconocido por un sistema de ese tipo puede percibirse como amenazante por una persona dadas las experiencias y tensiones personales previas de un contexto social particular (por ejemplo, el racismo que ha afectado a esa persona y a la comunidad), mientras que otra persona de otro trasfondo puede tener menos problemas con ello. En palabras de Matzner (2019): «las aplicaciones específicas de la IA se conectan de maneras muy diferentes con situaciones sociotécnicas preexistentes y las respectivas formas de subjetividad» (109). La tecnología de IA facilitará diferentes relaciones con diferentes subjetividades porque somos sujetos situados (118). En línea con la perspectiva de Foucault, esto implica que el poder de la IA no tiene que ver solo con manipulaciones, capacidades y sistemas (de arriba abajo); trata también de experiencias concretas y situadas y mecanismos de poder conformados por la tecnología. Se podría también decir (como lo haré al final de este capítulo): como seres vivientes, móviles y situados, *realizamos (perform)* el yo y el poder, y la IA desempeña un papel en estas realizaciones (*performances*), por ejemplo codirigiéndolas.

Continuaré desarrollando este marco para la reflexión sobre la IA y el poder centrándome en tres direcciones teóricas. Una primera fuente teórica, quizás obvia, para abordar, entender y evaluar el impacto de la IA sobre el poder es el marxismo. A continuación, elaboraré adicionalmente la idea de que la IA nos convierte en sujetos usando a Foucault, y desarrollaré la afirmación de que el poder se realiza (*is performed*) a través de la tecnología.

EL MARXISMO: LA IA COMO HERRAMIENTA PARA EL TECNOCAPITALISMO

Desde una perspectiva marxista, el poder de la IA está conceptualizado en términos de apoyo al capitalismo y a una clase social particular. Mediante la IA, nos gobiernan los gigantes tecnológicos y otros capitalistas. Estamos viviendo bajo una nueva forma de lo que Suarez-Villa (2009) denomina «tecnocapitalismo»: grandes corporaciones en su «búsqueda por el poder y el beneficio» (2) intentan controlar no solo todos los aspectos del dominio público, sino también nuestras vidas (ver de nuevo el argumento de Zuboff sobre el capitalismo de la vigilancia). Además, la IA se usa al servicio de Estados capitalistas y de sus agendas nacionalistas. Bartoletti (2020) compara la IA con el poder nuclear: la IA se usa para una nueva carrera internacional de armamento. Se podría añadir que la IA también es similar al poder nuclear en el sentido de que, de acuerdo al menos con este tipo de argumento, mediante la IA el poder es ejercido de arriba abajo desde una posición central y de una forma antidemocrática. De igual forma que nunca nos han preguntado si queremos el poder nuclear, nunca se nos ha preguntado si queríamos la vigilancia y la biométrica de la IA, los sistemas de decisión por IA, los procesamientos de datos de nuestro teléfono por parte de la IA, etc. De cara al ciudadano individual, la IA es un poder en el sentido de que permite el dominio y, para unos más que otros, la opresión. La economía de datos es política y es poderosa de cabo a rabo.

Sin embargo, no existe una cosa que sea una «IA» que nos oprima, como si la tecnología operase por su cuenta. La IA no debería entenderse como un factor aislado o un agente artificial atomístico. Siempre está conectada con los humanos, y la influencia de la IA sobre el poder ocurre siempre con y a través de los humanos. Si se pudiera decir de la IA que «tiene» poder (por ejemplo, poder sobre los humanos), ese poder lo es *a través* de los humanos y a través de la sociedad. Desde una perspectiva marxista, es la fuerza laboral viva, no las propias máquinas, lo que produce la plusvalía para los capitalistas (Harvey, 2019, 109). Además, la IA y la robotización buscan reemplazar el trabajo humano. Como señala Harvey (2019):

los robots no se quejan (excepto en la ciencia ficción), no replican, no exigen, no enferman, no trabajan más lentamente, no pierden la concentración, no se ponen en huelga, no piden aumento de sueldo, no se preocupan por las condiciones de trabajo, no quieren pausas para el café o simplemente no dejan de presentarse al trabajo (121-122).

Incluso el conocido como trabajo «inmaterial» (Lazzarato, 1996; ver también Hardt y Negri, 2000) que produce *software* o mundos virtuales requiere de humanos. Además, las elecciones políticas sobre la IA se hacen por Gobiernos y aquellos que la desarrollan y la emplean. La IA y la ciencia de datos son políticas y extremadamente poderosas en el sentido de que, en el contexto de una economía de datos, las decisiones son tomadas por las personas y sobre las personas a todos los niveles y en todas las fases de formas invisibles:

La elección sobre qué conjuntos de datos se estudian está tomada por personas. Es una decisión subjetiva, y política. Cada individuo, una vez que entra en un conjunto de datos, se convierte en parte de una nueva transacción entre ellos y la fuerza invisible que los ha puesto ahí, ha usado el conjunto de datos para entrenar un algoritmo y, en última instancia, ha tomado una decisión sobre ellos. Esto representa una asimetría de poder, y esta asimetría (el resultado de la elección y el poder) es lo que sostiene la política de datos y, en última instancia, la economía de datos. La economía de datos es política en todos los niveles, sobre todo porque algunas organizaciones poseen una gran cantidad de poder sobre otras al decidir quién entra en el conjunto de datos y quién queda fuera, una decisión que puede tener implicaciones de largo alcance (Bartoletti, 2020, 38).

Reconocer la política y el poder de la IA en términos de las «muchas manos» implicadas en estas operaciones no significa que no exista un uso centralizado, de arriba abajo, del poder. Tanto las grandes corporaciones como los Gobiernos usan la IA de una forma centralizada. Como hemos visto en el capítulo anterior, esto puede tomar la forma de la tecnocracia. Sætra (2020) defiende lo siguiente: priorizando el interés público, la IA puede lograr una forma de optimización racional, y la mayoría de problemas, entendidos correctamente, son «problemas técnicos susceptibles de ser tratados con la lógica del análisis y la optimización estadísticas» (2). Contra esta opinión se podría argumentar, en línea con la filosofía medioambiental y (añado) la filosofía de la tecnología, que las cuestiones técnicas son también políticas, y que los humanos, como animales políticos y agentes morales que pueden tener compasión y sabiduría, necesitan estar involucrados en estas cuestiones y tener responsabilidad respecto de ellas (5-7). Sin embargo, Sætra cree que, si la IA se desarrollara más, podría posiblemente resolver cuestiones difíciles para nosotros, ya que se desempeñaría mejor que los humanos a la hora de crear e identificar las mejores políticas, especialmente cuando se trata de áreas como la ciencia, la

ingeniería y ciertas cuestiones sociales y macroeconómicas complejas (5). Sin embargo, se podría objetar que estas cuestiones son también políticas, y que la política no se puede ni debe abordarse exclusivamente de una forma tecnocrática, ya que viola el principio de la democracia y dado que el juicio humano es necesario en política (ver el capítulo anterior).

Aun así, estos argumentos se podrían presentar sin tener en cuenta el capitalismo. Desde un punto de vista marxista, el problema principal no es tanto la tecnocracia y la falta de democracia como tales, sino un sistema socioeconómico específico con una lógica propia. Por ejemplo, Dyer-Witheford, Kjosen y Steinhoff (2019) observan que el capitalismo actual está «poseído» por la cuestión de la IA y argumentan que esta es un instrumento del capital y de la explotación. La IA no solo tiene una lógica tecnológica, sino también social, en particular la lógica de producir plusvalía. Contribuye a la creación y mantenimiento de un orden social particular: un orden capitalista (1-2). La IA reemplaza el trabajo y, si no lo hace, lo intensifica, y la amenaza de ese reemplazo ayuda a intimidar a los trabajadores. Las personas se convierten en algo prescindible o se les hace sentir que lo son (5). Mientras que algunos socialistas ven a la IA como una oportunidad para crear una sociedad diferente, por ejemplo mediante una renta básica universal, Dyer-Witheford, Kjosen y Steinhoff se centran en los problemas: nuevas formas de explotación y el amenazador panorama del capitalismo sin humanos.

Sin embargo, lo que afecta de la IA a los seres humanos no está limitado a la esfera de la producción y del trabajo, definidos de forma estricta. La IA se vuelve parte de la producción, pero también extrae conocimiento y conforma nuestra cognición y emociones. Desde una computación afectiva (Picard, 1997) a una IA afectiva, las tecnologías digitales intervienen a nivel personal, íntimo y emotivo. La «IA emotiva» (McStay, 2018) se usa para discernir estados emocionales, hacer análisis de sentimientos y medir la felicidad. Por ejemplo, las empresas pueden usar el análisis de sentimientos para reconocer, monitorizar y manipular los estados afectivos de las personas: esta es una forma de capitalismo cognitivo, pero también afectivo (Karppi *et al.*, 2016). Estas «operaciones que apuntan a tu personalidad, humor y emociones, tus mentiras y vulnerabilidades»

(Zuboff, 2019, 199) conducen a nuevas formas de explotación, dominación y manipulación política, por ejemplo haciendo campañas orientadas por los datos (ver Simon, 2019; Tufekci, 2018). Las redes sociales también favorecen mensajes emocionalmente comprometidos: el contagio afectivo (Sampson, 2012) se usa para influir en las masas. Puede incrementar el extremismo y el populismo, quizás incluso conducir hasta la violencia y la guerra.

La manipulación de las emociones por motivos políticos evoca viejas discusiones filosóficas sobre «las pasiones», «el afecto», «las emociones», y así sucesivamente, desde Spinoza hasta la filosofía actual y la ciencia cognitiva, incluyendo discusiones sobre el papel de las emociones en la política y cuestiones relacionadas tales como: ¿cuál es el papel del cuerpo en la política y qué es el «cuerpo político»? ¿Es, políticamente hablando, nuestra capacidad de ser afectados nuestro punto débil? Por ejemplo, Hardt (2015) ha argumentado que el poder de ser afectado no es necesariamente una debilidad y que somos sujetos no soberanos. Desde este punto de vista, quizás la ira podría jugar un papel positivo en la política. Nussbaum (2016), por el contrario, argumenta en contra de las personas que creen que preocuparse por la justicia necesita de la ira: es normativamente inapropiado (7) y necesitamos más bien generosidad e instituciones imparciales relacionadas con el bienestar. Una posición más positiva del rol de «las pasiones» en la política la ofrecen Farkas y Schou (2020), quienes, como hemos visto en el capítulo 4, argumentan que la política y la democracia no consisten solo en hechos, razones y evidencia, sino también en enfrentamientos de diferentes posiciones y en «el afecto, las emociones y los sentimientos» (7). Las emociones son, por tanto, necesarias en una versión dinámica de la democracia, como la imagina Mouffe (ver capítulo 4). Se podría discutir cuál es el papel de la IA frente a tales condiciones. Esto contrasta con la opinión de que los ciudadanos ordinarios son demasiado emocionales como para involucrarse en el debate político, y que necesitamos, en lugar de emociones, bien tecnocracia, o bien una educación racionalista de los ciudadanos. Otro tema controvertido que afecta a las emociones y a la política es la pertenencia: el sentido de la pertenencia es quizás importante, pero puede conducir al nacionalismo y al populismo; el

auge de ambos también está influenciado por la IA (ver el capítulo previo). La forma en la que la IA impacta sobre los aspectos emocionales de la política, el capitalismo y la democracia necesita una mayor investigación.

Cuando se trata de adoptar una perspectiva crítica sobre la IA y el poder, el «colonialismo de datos» (Couldry y Mejias, 2019) es otro término que se ha utilizado para expresar que la explotación de los humanos y de las vidas humanas mediante la IA es inaceptable desde una perspectiva teórica crítica. La apropiación de datos es entendida en este caso a través de la historia del colonialismo: exactamente igual que el colonialismo histórico se apropió de territorio y recursos para obtener beneficios, el colonialismo de datos explota a los seres humanos a través de la apropiación de datos (Couldry y Mejias, 2019). También se invoca al colonialismo en las discusiones sobre la IA y el sesgo, como hemos visto en el capítulo 3.

El *nudge* también puede ser visto como problemático desde la perspectiva de la teoría crítica: el hecho de que es una forma más sutil de manipulación no la convierte en menos explotadora, ya que se usa para obtener beneficio de una forma y en un contexto capitalista. Y como ya se indicó en el capítulo 2, el *nudge* es también problemático por otras razones: sortea la capacidad humana para la toma de decisiones y la capacidad de juicio autónomas. La IA incrementa las posibilidades del *nudge*. Yeung (2016) muestra que las técnicas de guiado algorítmico en las tomas de decisiones se usan para dar forma al contexto de elección en el que tiene lugar la toma de decisiones individual. Habla sobre *hypernudges* («hiperempujones»), ya que estos se actualizan continua y generalizadamente, lo que conlleva implicaciones preocupantes para la democracia y la prosperidad humanas. Pero el *nudge* no es solo problemático debido a su influencia directa sobre los seres humanos concretos: también implica una cierta forma de considerar a los seres humanos, una forma que facilita las relaciones manipulativas y explotadoras específicas que son tan preocupantes. Couldry y Mejias (2019) en su crítica de la idea de Floridi (2014) de que somos organismos informacionales o *inforgs* escriben que, cuando somos (nos han convertido en) *inforgs*, nos encontramos indefensos ante la manipulación y la modulación: «Los *inforgs*, una vez que se hayan rediseñado para estar

siempre abiertos a los flujos de datos, son las criaturas perfectas para que gobierne el *hypernudge* y, en consecuencia, para estar continuamente disponibles para la modulación» (158). Y si pensamos en tales *inforgs* como criaturas de ciencia ficción, entonces es bueno que seamos conscientes de que, como las emociones humanas y la heurística, la ciencia ficción también la usa el capitalismo para apoyar su sistema (Canavan, 2015; Eshun, 2003). O ¿puede también usarse para criticar las agendas y operaciones capitalistas, empoderar a los ciudadanos y sugerir el oponer resistencia?

La forma marxista de abordar y analizar los problemas relacionados con la IA y el poder nos conducen a la cuestión de si hay una forma de oponer resistencia, transformar o derrocar al capitalismo. Como pregunta general, esta cuestión ha sido discutida desde Marx, y como tema independiente que es, está más allá del alcance de este libro. Con respecto a la IA, merece la pena señalar que algunos estudiosos de la teoría crítica vislumbran la posibilidad de alinear a la IA con los ideales de la justicia social y del igualitarismo y contribuir a una renovación estructural, quitándoles poder a las instituciones hegemónicas (McQuillan, 2019, 170). Sin embargo, considerar la IA como un instrumento para la resistencia o la revolución podría enfrentarse a desafíos considerables, dado que esperanzas y afirmaciones similares ya se hicieron con respecto a internet. Como Castells (2001) y otros han mostrado, internet nació al principio en el complejo militar-industrial, más tarde los *hackers* lo vieron como un lugar de liberación, experimentación e, incluso, como un tipo de comunitarismo virtual (a pesar de que actualmente el libertarianismo parece haber triunfado en Silicon Valley). Internet tiene, en verdad, un cierto tipo de transparencia y promete una estructura más «horizontal» del poder. También ha reestructurado el trabajo y las clases sociales, al menos hasta cierto punto. Dyer-Witheford (2015) muestra que la gente que trabaja en la industria tecnológica no encaja bien en las clases sociales. La ética *hacker* parece tener la capacidad de empoderar y tiene potencial para la resistencia frente a la hegemonía capitalista. Por ejemplo, como escribe Castells (2001, 139), los *hackers* pueden desestabilizar las páginas web de las agencias gubernamentales o corporativas consideradas como opresivas o explotadoras. Pero hay también críticas: de forma general, el uso de la

tecnología de la información y los que la desarrollan son dóciles frente a las prioridades corporativas y militares (Dyer-Witheford, 2015, 62-63). Muchos temen que la IA también vaya por esa dirección: en lugar de una herramienta de transformación social, muy bien puede convertirse en una herramienta de opresión y explotación. La democratización de la IA es a veces defendida por los gigantes tecnológicos pero, al mismo tiempo, estos son reacios a restringir su poder y no aceptan la intervención externa (Sudmann, 2019, 25). Existe una tensión entre lo que proclaman y lo que realmente hacen. Mientras que la investigación en IA está dirigida a menudo por el idealismo, cuando está dentro de las grandes corporaciones lo que parece prevalecer es la competencia (Sudmann, 2019, 24). La IA y los algoritmos pueden contribuir a fomentar la desigualdad social, por ejemplo en los Estados Unidos (Noble, 2018). Finalmente, se ha defendido que los escenarios de ciencia ficción en los que interviene la IA pueden ayudar a mantener, en lugar de cuestionar, el sistema socioeconómico dominante: el capitalismo. Como señala Harvey (2019): «las nuevas tecnologías (como internet y las redes sociales) prometen un futuro socialista utópico, pero el capital se ha apropiado de ellas con nuevas formas y modos de explotación y acumulación en ausencia de otras formas de acción» (113). Obsérvese, sin embargo, que existen sistemas socioeconómicos, por ejemplo en Europa, en los que el capitalismo de IA se encuentra subordinado, al menos hasta cierto punto, a las normas éticas y políticas de las sociedades democráticas.

FOUCAULT: CÓMO LA IA NOS SOMETE Y NOS CONVIERTE EN SUJETOS

No todo el mundo está de acuerdo con la opinión marxista sobre el poder o con la opinión de que el poder es algo ejercido centralmente. En su influyente obra sobre este tema, Foucault defendía que el poder no consiste (simplemente) en una autoridad política soberana que ejerce el poder, sino que está implicado en todas las relaciones e instituciones sociales. Además, fue crítico en su teoría del poder con lo que vio como «economicismo» marxista (Foucault, 1980, 88): existe algo más que poder económico. A pesar de que es cierto que el poder sirve para «la reproducción de las

relaciones de producción», manteniendo así la dominación de clase, el poder también sirve para otras funciones y «pasa a través de canales mucho más finos» (72). Como ya expliqué en la introducción de este capítulo, el poder, de acuerdo con Foucault, impregna la sociedad al completo y llega a alcanzar a los sujetos individuales y a sus cuerpos en varios contextos y formas, que no están limitadas al dominio económico. Además, Foucault piensa que los yoes y los sujetos son *algo que se hace*, se hacen ellos mismos, y que esto es también una forma de poder. Examinemos estas opiniones con mayor profundidad.

Castigo y vigilancia

En primer lugar, Foucault defendió que el poder adopta la forma del castigo y la vigilancia. En *Vigilar y castigar* muestra que, bajo el poder disciplinario moderno, los individuos son usados como objetos e instrumentos (Foucault, 1977, 170): el cuerpo se vuelve dócil, obediente y útil. Basándose en este armazón se podría argumentar que, a través de redes sociales impulsadas por la IA, las tecnologías de vigilancia, etc., se crean cuerpos dóciles. La economía de la atención a las redes sociales nos convierte en máquinas que navegan y hacen clic en las páginas de internet, y a las personas se las somete a vigilancia en aeropuertos y otros entornos de control de fronteras. La IA también contribuye a la creación de un nuevo tipo de panóptico (Fuchs *et al.*, 2012). Diseñado por el filósofo inglés Jeremy Bentham en el siglo XVIII, el panóptico fue, inicialmente, un tipo de arquitectura para las prisiones con la forma de una torre de observación central colocada dentro de un círculo de celdas carcelarias. Desde la torre, los guardianes pueden ver a todos los prisioneros, pero los prisioneros no pueden ver el interior de la torre, lo que implica que nunca saben si están siendo vigilados. Como concepto disciplinario, el panóptico aplica la idea de que la gente se comporta como si estuvieran siendo observados, sin saber si este es el caso. Esta es una forma de control más sutil: es un tipo de autorregulación y una «tecnología política» (Downing, 2008, 82-83). Para Foucault (1980), el panóptico constituye una nueva forma en la que se ejerce el poder: fue «una invención tecnológica en el orden del poder» (71).

Foucault se refería al diseño de la prisión de Bentham, pero la IA del presente se puede entender como una contribución a todos los tipos de panópticos menos visibles, por ejemplo en el contexto de las redes sociales. Y Foucault ya había escrito sobre lo que ahora se denomina gobernanza sobre la base de la ciencia de datos. El panóptico incide también sobre la administración y los datos, conduciendo a lo que Foucault denomina «vigilancia integral»: métodos usados primero localmente, pero después (en los siglos XVIII y XIX) por el Estado, por ejemplo por la policía y la administración napoleónica:

La gente aprendió cómo establecer informes, sistemas de marcado y clasificación, la contabilidad integral de los registros individuales. [...] Pero la vigilancia permanente de un grupo de internos o pacientes era una cuestión diferente. Y, en un cierto momento, estos métodos comenzaron a generalizarse (Foucault, 1980, 71).

Actualmente, experimentamos una mayor generalización de tales métodos impulsados por la IA. La actual gobernanza por medio de los datos, o «gobernanza algorítmica» (Sætra, 2020, 4), conduce a lo que Foucault (1977) denominó una «sociedad disciplinaria», que funciona de forma análoga al diseño del panóptico de Bentham, pero que ahora invade todos los aspectos de la vida social: los efectos del poder disciplinario «no son ejercidos desde una única atalaya, sino que son móviles, multivalentes e internos al propio tejido de nuestra vida diaria» (Downing, 2008, 83). La IA y la ciencia de datos puede hacer esto, por ejemplo a través de las redes sociales y de los teléfonos inteligentes, que impregnan la vida social de todas las personas. Como ya he explicado, esto crea una cierta inquietud sobre la libertad como autonomía, la democracia y el capitalismo. Pero con Foucault podemos también comprenderlo en términos de pequeñas operaciones descentralizadas de poder. Por ejemplo, cuando usamos las redes sociales, las personas no son simplemente las víctimas pasivas de autoridades y corporaciones: otros usuarios individuales también ejercen formas de poder por la forma en que interactúan entre ellos y con la plataforma. Existen formas diferentes de vigilancia entre pares y lo que Albrechtslund (2008) denomina «autovigilancia» y vigilancia «participativa», que no viola necesariamente al usuario (a diferencia del *bullying*), pero que pueden ser lúdicas o incluso que tienen la capacidad de

empoderar en la medida en que capacitan a los usuarios para construir su identidad, socializar con extraños, mantener amistades y encontrar oportunidades para la acción. Este enfoque refleja una comprensión descentralizada y horizontal del poder, que también podría ser usada para entender la IA en términos de sus micromecanismos de poder.

Dicho esto, las formas centralistas y jerárquicas de poder aún existen. Por ejemplo, la IA se puede usar para apoyar versiones de gobernanza del «estado de emergencia» que, por ejemplo, en respuesta del terrorismo, «deciden sobre quién será detenido y quién no, quién podrá tener vida de nuevo fuera de la cárcel y quién no» (Butler, 2004, 62). En nombre de la seguridad o de acciones contra el terror, la IA puede ser parte de formas de «gobernanza algorítmica» (ver de nuevo Rouvroy, 2013) que pueden ser usadas para ejercer el poder del Estado de una forma que delega decisiones sobre quién se encuentra dentro y quién fuera de la comunidad política. Incluso en democracias liberales de la UE se usa cada vez más la IA para el control de fronteras. En el siglo XXI, las formas de gobernanza que los lectores de Foucault pensaban que pertenecían al pasado están de vuelta, mediadas y facilitadas ahora por la IA y la ciencia de datos. Para combatir la pandemia de la COVID-19 se usa la vigilancia policial y medidas tradicionales disciplinarias como el aislamiento y la cuarentena, pero ahora están autorizadas por la alta tecnología: la IA sirve de ayuda en la diagnosis a través de tecnologías médicas de la imagen y desarrolla medicinas y vacunas, pero también en la trazabilidad de los contactos entre los individuos, pronosticando la expansión del virus basándose en los datos disponibles y rastreando y monitorizando a través de teléfonos o pulseras inteligentes a los pacientes que están en cuarentena en sus hogares. En otras palabras: la IA ayuda a la vigilancia de una forma muy vertical, de arriba hacia abajo. La «única» cosa que era necesaria para que esto ocurriera fue una pandemia. La IA es una nueva herramienta biopolítica que permite nuevas formas de vigilancia e, incluso, una nueva biopolítica de matar o dejar morir, con drones y sistemas de triaje (Rivero, 2020). Sin embargo, siguiendo a Foucault, debemos subrayar que la mayoría de la vigilancia actual no procede de las autoridades políticas (Gran Hermano), sino que ocurre fuera de las instituciones políticas formales, se expande por toda la

sociedad: actualmente, la IA lo puede «ver» todo e, incluso, «oler» todo, la conocida como «odogilancia» (*odorveillance*) (Rieger, 2019, 145).

Si esto es cierto, ¿quién gana nuevo poder, además del Estado? Si el poder emana de centros nuevos y diferentes, ¿quiénes o qué son? Una respuesta es las grandes corporaciones, especialmente los gigantes tecnológicos. Véliz (2020) afirma que, recolectando nuestros datos, los gigantes tecnológicos y los actores políticos transforman el conocimiento en poder. Existe una asimetría de poder, ya que «ellos saben casi todo sobre nosotros» (86). La IA permite la vigilancia y la manipulación, pero la propia recolección de datos ya es problemática. Véliz distingue entre el poder duro y el blando de la tecnología. El poder duro es aquel en el que los datos se extraen incluso si nos resistimos, por ejemplo rechazando conceder el permiso (55). El poder blando, por el contrario, funciona de forma diferente, a menudo manipulativa: «nos fuerza a hacer algo en beneficio de otros con la excusa de que es por nuestro propio beneficio. Capta nuestra voluntad para algo que va en nuestra contra. Bajo la influencia del poder blando, participamos en un comportamiento que socava nuestros mejores intereses» (58-59). Véliz ofrece el ejemplo de la situación en la que vamos bajando por la pantalla de nuevas publicaciones de Facebook: nos enganchamos porque tememos que nos perdamos algo, y esto está hecho a propósito, ya que así se logra capturar nuestra atención en contra de nuestros mejores intereses (59). Esto se consigue mientras utilizamos ordenadores y móviles inteligentes, pero también se usan dispositivos como robots personales y asistentes digitales para ejercer este poder blando.

Conocimiento, poder y la creación y modelación de los sujetos y los yoes

En segundo lugar, la IA y la ciencia de datos no son solo herramientas poderosas en el sentido de que permiten el castigo y la vigilancia. También generan nuevo conocimiento y codefinen quiénes y qué somos. A lo largo de su obra posterior a *Vigilar y castigar* Foucault defendió no solo que el conocimiento es un instrumento de poder, sino que también el poder produce conocimiento y nuevos sujetos. Google se empodera gracias a nuestros datos, pero hay más: «Ese poder permite a Google decidir qué se

considera como conocimiento sobre ti a través del uso que hace de tus datos personales» (Véliz, 2020, 51-52). Así, las compañías tecnológicas no solo actúan sobre nosotros, sino que nos construyen en cuanto que sujetos humanos. Producen nuestros deseos (por ejemplo, el deseo de deslizarse por una página web) y nos convierten en seres diferentes con una forma diferente de estar en el mundo (Véliz, 2020, 52). Estas tecnologías son incluso utilizadas en las humanidades, lo que también conduce a nuevas formas de conocimiento y poder. A través del análisis del discurso algorítmico, por ejemplo, se crea el conocimiento de una forma no antropogénica, sorteando las intenciones y teorías humanas:

La minería de datos y de textos crea patrones y, en consecuencia, formas de conocimiento visibles que no se agotan necesariamente en cuestiones intencionales. Aquí, todo lo que la inteligencia humana considera, en su narcisismo científico, como su campo genuino de actividad (ordenando y clasificando cosas, identificando similaridades y creando genealogías) queda relegado a los algoritmos (Rieger, 2019, 144).

De acuerdo con Foucault (1980), la identidad individual es también el producto del poder: «el individuo, con su identidad y características, es el producto de una relación de poder ejercido sobre los cuerpos, las multiplicidades, los movimientos, los deseos, las fuerzas» (74). Esta producción de la identidad ocurre actualmente en las estructuras y procesos sociales más «horizontales» de las redes sociales, donde estamos bajo la vigilancia y el castigo no solo de los agentes jerárquicos del poder como los Gobiernos y las empresas, sino también de nuestros pares y, en última instancia, de nosotros mismos. Sin ser conscientes de ello, estamos trabajando sobre nuestros cuerpos y conformando nuestro yo e identidad. A medida que interactuamos con otros en las redes sociales y somos analizados y categorizados por la IA, no somos solo las víctimas de la gobernanza capitalista y del biopoder, sino también de la autodisciplina, la autocuantificación y autoproducimos nuestra subjetividad. Para Foucault, los cuerpos disciplinados pueden encontrarse en cualquier parte de la sociedad: los medios digitales y la IA son por completo partes de esa sociedad.

Este énfasis sobre los cuerpos, que continúa, por ejemplo, en el trabajo feminista de Butler, hace que sea muy concreto el enfoque de Foucault del

castigo y de la creación de sujetos, pero siempre conectado con el nivel social. Foucault (1980, 58) defendió que diferentes sociedades necesitan diferentes tipos de cuerpos: desde el siglo XVIII hasta mediados del siglo XX, los regímenes disciplinarios de escuelas, hospitales, industrias, familias, etc., implicaban investir el cuerpo con una gran carga de poder, mientras que, posteriormente, se han ejercido sobre el cuerpo formas de poder más sutiles. Actualmente, nos podemos preguntar qué tipos de cuerpos se necesitan y se crean a través del castigo por parte de la IA y otras tecnologías digitales, y qué nuevas formas de poder sutiles y menos sutiles facilitan estas tecnologías. Parece que la «sociedad de la IA» contemporánea necesita cuerpos que puedan ser *datificados*, cuantificados e (in)movilizados para que tales datos y números se entreguen a través de interacciones con teléfonos inteligentes y otros aparatos. Esto constituye una forma de ejercer el poder sobre nuestros cuerpos más blanda y menos visible, pero no por ello menos invasiva. Los efectos que la IA tiene en nosotros en términos de castigo, poder y subjetividad no es entonces una cuestión de operaciones meramente «mentales» o de cognición, si la «cognición» la entendemos como descorporeizada y completamente inmaterial; también tiene efectos corpóreos. Pero tomando en consideración las lecciones de la ciencia cognitiva contemporánea (por ejemplo, Varela, Thompson y Rosch, 1991), cualquier cognición de la que merezca la pena hablar es cognición corporeizada. Nuestro pensamiento y experiencia dependen de nuestro cuerpo. El cuerpo desempeña un papel activo en los procesos cognitivos. Además, si tenemos en cuenta las intuiciones del poshumanismo (por ejemplo, Haraway, ver el capítulo siguiente), el cuerpo no es solo el cuerpo biológico: también puede entenderse como ligado y fusionado con lo material, como algo que tiene una naturaleza «ciborg». En este sentido, la afirmación, hecha por algunos académicos marxistas, de que el trabajo inmaterial está terminado es engañosa: lo que hacemos con nuestros cuerpos y mentes está muy conectado con las tecnologías materiales que usamos y tiene muchas consecuencias físicas, incluyendo efectos sobre nuestros cuerpos, nuestra salud y bienestar. Por ejemplo, cómo nos autodisciplinamos, somos castigados y producimos nuestro yo y nuestra subjetividad a través del uso de teléfonos inteligentes y sus *apps*

que operan con IA es algo que tiene efectos sobre nuestros músculos, ojos, etc., y puede conducir al estrés, emociones negativas, desórdenes del sueño, depresión y adicción. La IA, en este caso, puede muy bien ser «virtual» o «inmaterial», en el sentido de que se presenta en la forma de *software* y bases de datos, pero su uso y efectos son materiales y físicos, involucran el cuerpo y la mente. Usando la metáfora del vampiro de Marx: el capitalismo de la vigilancia *succiona ese trabajo vivo*.

Sin embargo, el poder y la producción de sujetos, cuerpos y conocimiento no es necesariamente una cosa mala por sí misma y, en muchos casos, no es necesariamente violenta ni restrictiva. Foucault argumentó que el poder y la fuerza también se pueden ejercer de formas sutiles que son físicas, pero no necesariamente violentas (Hoffmann, 2014, 58). Todo depende de cómo se haga y de cuáles sean los resultados. ¿Qué produce el poder y qué fuerzas crea la tecnología?

Desarrollaré ahora el enfoque productivo del poder de Foucault. La idea del moldeado del yo por el poder y lo tecnológico también se puede basar en el trabajo posterior de Foucault, en el que escribe sobre las transformaciones del yo en la antigua Grecia y en la cristiandad que se logran mediante la aplicación de «tecnologías del yo»: una de las formas en las que los seres humanos desarrollan conocimiento sobre sí mismos.

Las tecnologías del yo permiten que los individuos se vean afectados por sus propios medios, o con ayuda de otros, por cierto número de operaciones sobre sus propios cuerpos y almas, pensamientos, conducta y forma de ser, para transformarse y obtener un cierto estado de felicidad, pureza, sabiduría, perfección o inmortalidad (Foucault, 1988, 18).

Foucault no considera estas «tecnologías» como materiales, y las distingue de las tecnologías de producción. Está interesado en la «hermenéutica del sujeto» y en la virtud y práctica del autocuidado en la filosofía grecorromana y en la espiritualidad y prácticas cristianas (19). Sin embargo, revisando a Foucault bajo la luz de la filosofía de la tecnología contemporánea, se podría considerar la escritura como una tecnología material del yo que ayuda al autocuidado, autoconstitución y, ciertamente, a alcanzar la virtud. Cuando los filósofos antiguos, los monjes cristianos y los académicos humanistas escribieron sobre sí mismos, ejercitaban tanto cuidados como poder sobre sí mismos. Se podría decir que tecnologías

como la IA, que permiten la autovigilancia, autorrastreo, autocuidado y autocastigo, también se usan como «tecnologías del yo»: no simplemente para la dominación y el castigo de otros (a pesar de que este aún podría ser el caso, considérese de nuevo el análisis marxista y el anterior de Foucault), sino también para ejercer una forma de *poder sobre uno mismo*. Considérense las aplicaciones para cuidar de la salud, cuyo objetivo es disciplinar la alimentación y el ejercicio físico, o las *apps* para la meditación: estas se usan para el autocuidado pero, al mismo tiempo, implican ejercicio del poder sobre uno mismo, sobre el alma y el cuerpo, que conduce a un tipo específico de autoconocimiento (por ejemplo, cuantificación del yo), implica la operación de fuerzas físicas y constituye un tipo particular de sujeto y de cuerpo. En lugar de restringir algo preexistente, el poder es aquí productivo, ya que lleva a algo (un yo, un sujeto, un cuerpo) a la existencia. En este sentido, es habilitante en lugar de limitante. La IA se puede utilizar dentro de tales prácticas de autoconstitución. Se deberían plantear entonces preguntas críticas no solo sobre qué tipo de yo y sujeto produce la IA, sino también sobre qué tipo de autocuidado y práctica del yo están implicadas. La tecnología misma no tiene ese poder: es la tecnología en cuanto usada dentro de una práctica del autocuidado, autocastigo, etc. Pero la tecnología influye en el tipo particular de automoldeado. Por ejemplo, se podría afirmar que la IA y la ciencia de datos producen un «yo cuantificado» a través de una práctica de autorrastreo, que produce también un tipo particular de conocimiento: un conocimiento que toma la forma de números.

Esta creación del yo puede también teorizarse usando la obra de Judith Butler. Como Foucault, Butler considera el poder como algo productivo, pero para ella toma la forma de la constitución *performativa* del yo. Usando la descripción de Austin (1962) sobre cómo algunos actos de habla hacen cosas, argumenta que nuestros yoes e identidades (por ejemplo, identidades de género) no son esencias, sino que se encuentran *performativamente* constituidos (Butler, 1988); el género es un tipo de resultado *performativo* (Butler, 1999). Esta consideración deja a los yoes y a las identidades tanto sin fijar como sin asumir (Loizidou, 2007, 37). En línea con Foucault, argumenta que no se trata solo de sometimiento (por ejemplo, como

castigo), sino también de convertirse en un sujeto (por ejemplo, subjetivación). Pero al subrayar la dimensión *performativa* de esta práctica y operación, afirma que su análisis del poder es menos pasivo que el de Foucault (Butler, 1989). El poder tiene que ver con los actos que constituyen al sujeto. Esto no consiste simplemente en que otros nos conviertan en algo: nosotros también nos moldeamos a nosotros mismos, por ejemplo diciendo cosas. Y esta es también una cuestión temporal: Butler (1993) entiende la *performatividad* no como un acto, sino como una práctica reiterativa: «la *performatividad* debe ser entendida no como un “acto” singular o deliberativo, sino, más bien, como la práctica reiterativa mediante la cual el discurso produce los efectos que nombra» (2). Esto se podría reconciliar con Foucault, quien habla sobre las prácticas del autocuidado. E, inspirado por Bourdieu (1990), se podría añadir: la autoconstitución es una cuestión de *habitus*. La constitución del yo opera a través de la operación habitual y *performativa* de fuerzas sobre uno mismo.

Con todo, la concepción de la *performatividad* de Butler y su concepción de la política (Butler, 1997) continúan centradas en el lenguaje. Como en Foucault, el énfasis se pone en el discurso. A esto debemos añadir la idea de que el yo, la identidad y el género se producen y realizan (*perform*) no solo a través del lenguaje, sino también a través de las prácticas tecnológicas. Junto con la escritura, existen otras prácticas tecnológicas: tecnologías web 2.0 (Bakardjieva y Gaden, 2011) tales como las redes sociales, pero también la IA. La constitución del yo a través de la IA bien puede tener aspectos lingüísticos, pero, como he dicho, también es profundamente tecnológica y material, como cuando se produce un «yo cuantificado». Además, siempre es también una cuestión social. Tanto el yo como el otro toman forma en el proceso. Por ejemplo, a través del uso de aplicaciones para correr y otras tecnologías portátiles de seguimiento automático, el otro se convierte en un objeto de examen y competición (Gabriels y Coeckelbergh, 2019).

Estas formas tecnológicas de automoldearse y autocuidarse plantean al menos dos problemas. Primero, la cuantificación del yo y del otro sugieren equivocadamente que el yo o el otro pueden ser reducidos a una recopilación de información digital: esto es, que el yo digital es el yo real (véase de nuevo *datificación*) o, al menos igual de problemático que lo

anterior, que el yo o el otro digital es *más* yo que el yo o el otro no digital. Esta última suposición parece estar operativa al menos en la fantasía transhumanista de la inmortalidad y la resurrección a través de la subida de archivos a la web. Kurzweil ha imaginado que el aprendizaje automático permitirá reconstruir una versión digital de su propio padre ya fallecido, haciendo posible hablar con un avatar suyo: «será tan realista que será como hablar con mi padre». En realidad «se parecería más a mi padre que mi propio padre, si viviera» (Berman, 2011). Criticando esta concepción, Andrejevic (2020) ha argumentado que Kurzweil busca crear una imagen idealizada del yo (y del otro) que sea más coherente y consistente que un sujeto real. Pero un sujeto real siempre está «constituido por sus lagunas e inconsistencias», de ahí que cualquier intento de perfeccionar el yo y el sujeto «equivalga a intentar eliminarlo» (1). La automatización de la producción mental y de la inteligencia humana relevante que sea necesaria para esa empresa intenta también reconfigurar y, por lo tanto, eliminar al sujeto: al abstraer de las tareas del sujeto sus motivaciones, intenciones y deseos, se deja de lado «el nivel reflexivo de la subjetividad» (5). Esto implica no solo que se eluda el juicio y el pensamiento humanos, como Arendt aceptaría, sino también que el sujeto humano no es solo superfluo, sino incluso un estorbo. Al igual que la fantasía de un sujeto automatizado se ve amenazada por la realidad de los sujetos reales que «pueden ser impredecibles, recalcitrantes y, por lo demás, irracionales de formas que amenazan a los sistemas de control, la gestión y el gobierno» (2) y obstaculizan el funcionamiento fluido y sin fricción de la sociedad automatizada, la fantasía de cargar el yo en la web intenta crear un avatar digital que *dejaría de ser en absoluto un yo o un sujeto*. Este es también un problema cuando intentamos moldearnos a nosotros mismos y a otras personas con la ayuda de las redes sociales y de la IA: intentamos moldearnos como *algo*, quizás avatares, pero no personas, yoes o sujetos. Además, como señalan Couldry y Mejias (2019, 171) utilizando a Hegel, estar mediado no es un problema en sí mismo, pero una vida en la que falte la relación reflexiva con uno mismo, sin espacio para estar consigo mismo, no es una vida libre.

Además, todo este automoldeado es agotador y potencialmente explotador, y la autoexplotación es también problemática. Hemos pasado de la sociedad disciplinaria de Foucault a lo que Han (2015) denomina la «sociedad del logro» (8). En el capítulo 3 ya vimos que el capitalismo produce yoes ansiosos que han internalizado el imperativo de rendir, y que tienen miedo de ser reemplazados por máquinas. Han argumenta que, en la sociedad contemporánea, las prohibiciones y obligaciones se encuentran sustituidas por «proyectos, iniciativas y motivación» (9). Mientras que la sociedad disciplinaria produce lunáticos y criminales, «la sociedad del logro crea depresivos y perdedores» (9). Los individuos depresivos están cansados de tener que llegar a ser ellos mismos. Los individuos se explotan a sí mismos, especialmente en contextos laborales en los que tienen que alcanzar logros y rendir. Se convierten en máquinas. Pero parece que la resistencia es imposible en este caso, ya que el explotador y el explotado son el mismo: «el exceso de trabajo y de rendimiento se convierte en autoexplotación» (11). La depresión «surge en el momento en el que el sujeto-logro *deja de ser capaz de ser capaz [nicht mehr können kann]*» (10, énfasis de Han en alemán). Esto podría relacionarse con un análisis marxista: el sistema capitalista exige esta autoexplotación. Desde el punto de vista del capitalista es un sistema brillante, ya que parece que las personas solo pueden culparse a sí mismas cuando no rinden o no logran lo suficiente, así como cuando fracasan en trabajar lo suficiente sobre sí mismas. Incluso en la esfera privada, nos sentimos constantemente como si tuviéramos que realizar este autotrabajo. La IA y las tecnologías relacionadas se usan para incrementar nuestro desempeño laboral, pero también las usamos para trabajar sobre nosotros mismos, hasta que *ya no podemos más [nicht mehr können]*. Incluso la autoconstitución se ha convertido en una cuestión de logro, promovido por las tecnologías que nos monitorizan, analizan e impulsan nuestra performatividad hasta que ya no podemos rendir más y nos quemamos. Oponer resistencia a tal sistema de poder y gobernanza es difícil, ya que parece que solo nos podemos culpar a nosotros mismos por no dar la talla. Deberíamos haber usado el tipo de *app* correcta y haber hecho más autotrabajo. Si estamos deprimidos o quemados, es nuestra culpa, y un fracaso de cara al logro.

En principio, sin embargo, son posibles diferentes matices y tecnologías del yo. El marco teórico de Foucault deja abierta la posibilidad de que el uso productivo del conocimiento y autoconstituyente del poder también pueda tomar otras formas de empoderamiento. O, por ponerlo en términos de *performatividad*: son posibles otras *performatividades* o realizaciones del yo. Ofreceré ahora ciertas sugerencias que se encuentran en mi propia obra sobre *performatividad* y tecnología para mostrar las posibilidades de una visión del poder orientado hacia la *performatividad* para comprender y evaluar la IA.

«TECNOPERFORMANCES», PODER E IA

Como hemos visto, Butler usa el término *performance* para conceptualizar la constitución del yo, pero también podemos utilizarlo para conceptualizar el uso de la tecnología, subrayar la forma en que la tecnología tiene poder sobre nosotros y revelar conexiones entre las *performances* de los individuos y su contexto político (Coeckelbergh, 2019b; 2019c). Aquí, el objetivo de conectar la tecnología con la *performance* no consiste simplemente en decir que la tecnología se usa en *performances* artísticas (Dixon, 2007), sino que una *performance* se puede utilizar como metáfora y concepto con el que reflexionar sobre la tecnología. Con respecto al poder, este enfoque nos permite describir y evaluar lo que ocurre en términos de poder cuando las tecnologías tienen una mayor agencia, por ejemplo he argumentado que nos *dirigen* y *coreografían* (Coeckelbergh, 2019b). Cuando participamos en *tecnoperformances* (Coeckelbergh, 2019c), la tecnología adquiere un papel cada vez más importante a la hora de liderar y organizar. Hay aquí un sentido en el que no solo *performamos* con la tecnología: «la tecnología también *performa* con nosotros». Los humanos no se encuentran ausentes: *co-performamos*, *co-dirigimos*, *co-coreografiamos*. Pero la tecnología también conforma la *performance*. La cuestión es, entonces, qué obra, coreografía, etc., queremos crear con la tecnología y cuál es nuestro papel en estas *performances* (Coeckelbergh, 2019b, 155). Basándonos en este enfoque, se podría insistir, de nuevo, en que la IA no es solo un instrumento

usado por los humanos para ejercer el poder, sino que tiene también efectos imprevistos, y que uno de estos efectos se puede describir como sigue: cómo y a medida que la IA alcance una mayor agencia, por ejemplo en la forma de coches autónomos, robots y algoritmos que operan en internet, pero también a medida en que la influencia imprevista de la IA se vuelva más generalizada, la IA se convierte en coreógrafa, directora, etc., de nuestros movimientos, discursos, emociones y vida social. No es simplemente una herramienta o una cosa, sino que organiza la forma en que hacemos las cosas. De nuevo, esto no implica que los humanos no estemos implicados en ello, o que no seamos responsables, sino que la IA tiene un papel mayor que el simplemente instrumental en el que conforma lo que hacemos y cómo lo hacemos. Tiene el poder de organizar nuestras *performances* y cambia el campo de fuerzas y relaciones de poder.

El uso del término *performance* en relación con la IA también incorpora diferentes dimensiones del uso de la tecnología que se relacionan con el poder. En primer lugar, nos permite decir que el uso de la IA es siempre una cuestión social, entendida como *co-performance*. Involucra a seres humanos que interactúan y operan en un contexto social y, por tanto, también en un contexto político. Puede implicar también a una «audiencia» que responda a su uso. Se podría decir que la IA está siempre situada en este entorno social que, como muestra Foucault, está impregnado por el poder. Por ejemplo, el uso de la IA por los gigantes tecnológicos tiene lugar dentro de un contexto sociopolítico y tiene una audiencia de usuarios y ciudadanos que responden a lo que hacen estas compañías. Esta «audiencia» también tiene poder y es parte de las relaciones de poder. En segundo lugar, si el uso de la IA se conceptualiza como *performance*, esto también significa que el cuerpo queda involucrado. Como señalé anteriormente, el hecho de que la IA se presente de una forma «inmaterial» o «virtual», por ejemplo en términos de *software*, no implica que el efecto del poder no se refleje en los cuerpos. Las *tecnoperformances*, como todas las *performances*, implican cuerpos humanos. Esto está en línea con la atención al cuerpo que prestan Foucault y Butler, pero no ligan exclusivamente el concepto del cuerpo al discurso, al conocimiento y a la identidad. La forma en que el poder opera aquí es también de manera

completamente literal, sobre el *movimiento*, moviendo cuerpos. Cuando uso la IA a través de una aplicación de mi teléfono inteligente, no soy un usuario descorporeizado que usa solamente funciones «mentales» o «cognitivas»: estoy moviendo mi cuerpo y mis manos, mi cuerpo está parcialmente inmovilizado, etc. Esto es así porque la IA y sus diseñadores coreografían los movimientos que son necesarios para operar el dispositivo y la aplicación de cierta manera y, por lo tanto, ejercen el poder sobre mí y sobre mi cuerpo. En tercer lugar, la noción de *performance* también opera en el aspecto temporal. El ejercicio del tecnopoder por y a través de la IA tiene lugar en el tiempo e, incluso, lo configura, en el sentido de que da forma a nuestra experiencia del mismo y configura nuestras historias, días y vidas. Por ejemplo, cogemos regularmente nuestro teléfono y miramos mensajes y recomendaciones: esto se convierte en parte de nuestra rutina diaria. En este sentido, la IA tiene el poder de definir mi tiempo. Y mediante la recolección y análisis de datos, mi historia queda configurada en términos de las categorías estadísticas y los perfiles elaborados por la IA. Esto ocurre no solo a nivel individual, sino también a nivel cultural y social: nuestro tiempo se convierte en el tiempo de la IA y la IA moldea la narrativa de nuestras sociedades.

Este enfoque se hace eco del pensamiento foucaultiano, incluyendo la teoría foucaultiana de la danza y de la *performance*. Kozel (2007) se ha referido al poder y al conocimiento para hacer afirmaciones sobre lo que McKenzie (2001, 19) denomina «los mecanismos del poder *performativo*». La *performance* se contempla como «distribuida a lo largo de temporalidades, redes y cuerpos de todo tipo» (Kozel, 2007, 70). Y así, se podría añadir, es poder: el poder en y de las *tecnoperformances* está también distribuido a través de temporalidades, redes y cuerpos. Además, nos encontramos de nuevo con la concepción «productiva» del poder: la IA tiene poder para conformar nuestro tiempo (de una forma particular). En línea con Foucault y con Butler (1988), se podría argumentar que las *tecnoperformances* que incluyen IA no solo nos castigan y nos ponen bajo vigilancia, sino que también nos constituyen como nuevos sujetos, ciudadanos e identidades. También producen un tipo particular de yo y de subjetividad. Me llego a entender a mí mismo de una forma particular a

través del uso de la IA. Por ejemplo, a través de estas *tecnoperformances*, podemos llegar a tener un sentido del yo conectado (Papacharissi, 2011) o, como ya he sugerido anteriormente, un sentido del yo *datificado*³. Cheney-Lippold (2017) argumenta que los algoritmos y las corporaciones que los emplean, como Google y Facebook, usan los datos para construir nuestros mundos e identidades. En este sentido, y como reza el título de Cheney-Lippold, «somos datos», y cada vez lo creemos más. Desde la perspectiva de una teoría crítica, se debería señalar que esto es increíblemente útil para las empresas cuyo modelo de negocio se basa en monetizar nuestros datos, como Facebook. No somos solo consumidores, sino que, al mismo tiempo, somos productores de datos: trabajamos para estas empresas, que nos explotan mediante lo que Fuchs y coautores (2012, 57) denominan «una máquina de clasificación panóptica», que identifica los intereses de los usuarios, los clasifica y, a continuación, está en condiciones de realizar publicidad dirigida. Pero añado que esta construcción del mundo y del yo (y la explotación relacionada con él) no es simplemente algo que nos ocurra a nosotros. Nuestras *tecnoperformances* con la IA se desarrollan a medida que nos involucramos con la tecnología: es un proceso activo y es el resultado del esfuerzo humano. No es simplemente que otros, o los algoritmos de la IA, nos conviertan en datos. A través de nuestras *performances* con la tecnología *también nos convertimos en datos* en la medida en que constituimos *tecnoperformativamente* nuestros yoes en las redes sociales y en otras partes. Así, contribuimos tanto a nuestra autoconstitución como a nuestra explotación.

Considerar la tecnología en términos de (tecno)*performances* está en línea con las concepciones de la tecnología que la consideran una actividad. Tales concepciones nos permiten incorporar las dimensiones sociales y políticas (Lyon, 1994). La tecnología puede tratar de artefactos y de cosas, por supuesto, pero para estudiar su dimensión política necesitamos mirar a lo que hacemos con la tecnología y lo que esta nos hace, y cómo ambas cuestiones se encuentran incrustadas en un contexto social (y de conocimiento). Considerar la tecnología como una actividad y como una *performance* también nos permite subrayar de nuevo que los humanos están siempre involucrados en el proceso. Mientras que las orientaciones

contemporáneas de la filosofía de la tecnología como la posfenomenología proclaman correctamente que las tecnologías «hacen» cosas (Verbeek, 2015), en el sentido de que comoldean la percepción y la acción humanas (por ejemplo, un microondas conforma nuestros hábitos alimenticios, la tecnología de ultrasonidos conforma cómo experimentamos el embarazo, etc.), las actividades de la tecnología siempre incluyen humanos. Volveré sobre esta cuestión cuando considere las opiniones poshumanistas en el siguiente capítulo.

Finalmente, dado que los humanos continúan codirigiendo y comoldeando sus *performances* sociales, corporales y temporales y, así, participan en el ejercicio y circulación del poder, debemos preguntarnos *qué* humanos (co-)coreografían nuestras *tecnoperformances*, y si la participación en *ellas* es siempre voluntaria. Como se pregunta Parviainen (2010): ¿quién nos coreografía? Por ejemplo, se podría afirmar que los gigantes tecnológicos nos coreografían mediante el uso de la IA integrada en nuestras aplicaciones y dispositivos, que diseña cada vez más qué hacemos y que da forma a nuestras historias, pero, ya que normalmente no somos conscientes de este hecho, y dado que las tecnologías están diseñadas para ser persuasivas, nuestra participación en estas *performances* e historias difícilmente pueden considerarse como voluntarias. Y si las *tecnoperformances* se encuentran siempre ligadas a un contexto social y político más amplio, es importante preguntarnos a quién se *permite* participar en las *performances* de la IA (tanto en el uso como en el desarrollo), esto es, quién está incluido y quién excluido, y cuáles son las condiciones de esta participación.

Primero, a muchos de nosotros no solo se nos facilita una pseudoelección (Bietti, 2020) cuando se trata de consentir el uso, sino que también estamos excluidos cuando se trata de moldear el desarrollo de la IA y cómo se usa. En este punto, estamos en manos de los gigantes tecnológicos. Incluso los Gobiernos se arreglan a menudo con lo que ofrece el sector tecnológico: en muchos países, la regulación es mínima. Una perspectiva orientada hacia la *performance* nos permite hacernos preguntas críticas sobre cuestiones como estas: ¿quiénes son los actores y los coreógrafos en la *performance*? ¿Quién está excluido de actuar y

coreografiar? ¿Qué actores y coreógrafos tienen más poder que otros? Y ¿podemos desarrollar estrategias para escapar del control absoluto? Estas cuestiones se relacionan, de nuevo, con la discusión sobre la democracia.

Segundo, las *performances* de la IA pueden también ser altamente políticas en el sentido de que se encuentran ligadas a las políticas gubernamentales. Usando el concepto de coreografía, Parviainen y yo mismo hemos defendido que la IA y la robótica se usan en el contexto del interés y de las estrategias políticas (Parviainen y Coeckelbergh, 2020). Mostramos que las *performances* de Sophia, un robot humanoide del que se afirma que usa IA, están ligadas con políticas de la IA: «Las *performances* de Sofía no solo sirven al interés de una empresa privada (Hanson Robotics); también sirvieron a los intereses de aquellos que buscaban expandir las tecnologías involucradas y los mercados relevantes conectados a estas tecnologías» (7). Usando el término «poder», se podría también abordar este tema como sigue: el sector privado involucrado en el desarrollo de la IA y de la robótica organiza *tecnoperformances* como una forma de expandir mercados e incrementar así su poder y beneficio. De forma similar, los Gobiernos pueden apoyar tales *performances* y participar en ellas para llevar a cabo sus planes y estrategias con respecto a la IA, incrementando en consecuencia su poder en respuesta a los competidores, esto es, otras naciones. Además, las discusiones sobre el estatus de la inteligencia o la ética de los robots sociales pueden enmascarar su dimensión política, ya que hace parecer como si solo se pudieran plantear cuestiones técnicas o éticas sobre la IA y la robótica: cuestiones relacionadas con la interacción inmediata y el entorno de la tecnología, pero no sobre los campos más amplios sociales y políticos. Esto sugiere de forma engañosa que estas tecnologías son neutrales respecto del poder y la tecnología. Las *performances* con IA y las discusiones relacionadas pueden ocultar de este modo que, como sucede con todas las tecnologías, la IA es y puede ser muy política y poderosa. Los investigadores y periodistas pueden revelar este contexto político más amplio, relacionando *performances* de la IA locales y concretas con lo que ocurre al nivel «macro» de la política y sus juegos de poder.

Este uso del término *performance* y sus relaciones con el poder ofrece entonces un marco para adoptar una perspectiva crítica sobre la IA, que es compatible con, y puede estar apoyada por, tanto los análisis marxistas como los enfoques foucaultianos. Ayuda a revelar los miles de formas en las que la IA está conectada con el poder: el poder ejercido en y a través de *tecnoperformances* y que circula entre estas *performances* y un campo de poder y de sus actores más amplio, como las corporaciones y los Gobiernos.

CONCLUSIÓN Y PREGUNTAS PENDIENTES

Este capítulo, que ha mostrado cómo hablar sobre el poder y la IA, nos permite introducir una teoría social y política que ayuda a conceptualizar los aspectos políticos de la IA. Naturalmente, este ejercicio no trataba solo sobre el poder, sino que estaba también relacionado con otros conceptos y cuestiones políticas. Por ejemplo, la discusión sobre el *nudge* se relaciona con problemas en torno a la libertad y la cuestión del sesgo, que ya figuraba en el capítulo 3 sobre la igualdad y la justicia. Se podría decir más sobre cualquiera de estos vínculos; por ejemplo, la cuestión sobre el sesgo y la IA (Bozdag, 2013; Criado Perez, 2019; Granka, 2010) se puede abordar como una cuestión de poder, pero también como de justicia y de igualdad: si existe un sesgo en el filtro algorítmico, en los algoritmos de búsqueda o en los conjuntos de datos con los que se entrena la IA, entonces esto tiene que estar relacionado con las personas que ejercen el poder sobre otras. También se podría decir que existe una estructura de poder particular (por ejemplo, capitalista, patriarcal, etc.) en una sociedad particular, que conduce al sesgo a través del uso de la IA. Parece que tener una pluralidad de conceptos es útil para desentrañar estas cuestiones. Sin embargo, el concepto de poder ha sido una lente útil a través de la cual hemos podido analizar y discutir sobre la política de la IA. El capítulo ofrece aún otra forma única de conceptualizar que la tecnología es política y *cómo* lo es. Muestra cómo hablar sobre el poder nos ayuda a analizar lo que exactamente está ocurriendo y qué podría haber de problemático en ello.

Reunir conceptualmente a la tecnología y a la política y, especialmente, afirmar que la IA no tiene efectos no instrumentales sobre el poder continúa

siendo problemático desde un punto de vista moderno, ya que la tecnología y la política se encuentran, en la modernidad, en esferas separadas. La primera se supone que trata sobre la técnica y las cosas materiales, mientras que la política se supone que lo hace sobre los humanos y la sociedad. Esta idea moderna hunde sus raíces en la filosofía antigua, al menos desde Aristóteles, y continúa ejerciendo hoy en día su influencia: se supone que la IA es políticamente neutral, y la política se supone que trata sobre los fines para los cuales los humanos usan la IA. Las discusiones de este capítulo cruzan esta línea de división moderna hablando, por ejemplo, sobre cómo la IA moldea nuestro yo creando nuevas formas de subjetividad y cómo nos coreografía. En las poderosas *performances* de IA se mezclan los fines y los medios y, en última instancia, también lo hacen los seres humanos y la tecnología. Sin embargo, como he subrayado repetidamente, la idea no es que los humanos hayan sido reemplazados por la tecnología o que la tecnología hace todo eso por sí sola. La cuestión importante en la aplicación de términos tales como «coreografía» a la IA no es afirmar que los humanos no están involucrados en el ejercicio del poder y en las *tecnoperformances*, sino que a la IA se le otorga a veces una mayor agencia y, a través de efectos imprevistos, co-moldea cómo hacemos las cosas y quiénes/qué somos. En este sentido, la IA tiene poder sobre nosotros.

Al desdibujar la línea entre la tecnología y la política, en particular entre la IA y el poder, podemos hablar de la IA como «poder artificial»: no porque la IA sea *todopoderosa*, sino porque el poder se ejerce a través de la IA. La IA es muy poderosa y también política solo como parte de las *tecnoperformances* humanas que moldean lo que hacemos y quiénes/qué somos.

Con todo, existe otro límite relacionado que merece discusión: el límite entre humano y no humano. A menudo se acepta que la política, y, por tanto, también la política de la IA, trata solo sobre seres humanos, pero esto se puede discutir y, de hecho, ha sido discutido. Este es el tema del siguiente capítulo.

³ La denominada «datificación» es la consideración de un proceso o actividad que era previamente invisible y convertirlo en datos.

CAPÍTULO 6

¿Qué sucede con los no humanos? Política medioambiental y poshumanismo

INTRODUCCIÓN: MÁS ALLÁ DE UNA POLÍTICA ANTROPOCÉNTRICA DE LA IA Y LA ROBÓTICA

La mayoría de teorías discutidas en los capítulos previos aceptan que la filosofía política y que, por lo tanto, también la aplicación de la filosofía política a la IA, trata de la política para los seres humanos. Los principios políticos como la libertad, la justicia, la igualdad y la democracia se supone que se refieren a la libertad para los humanos, la justicia para los humanos, y así sucesivamente. El *demos*, el público y el cuerpo político se acepta que consisten en humanos y en sus instituciones. Y la mayoría de la gente cree que el término «poder» es solo aplicable a las relaciones humano-humano. Si el poder circula a través del cuerpo social, como argumenta Foucault, entonces este cuerpo se concibe como si consistiera exclusivamente en humanos. Además, la ética y la política de la IA y de la robótica se abordan a menudo de una forma antropocéntrica: se afirma que la IA y la robótica deberían ser antropocéntricas, en cuanto que opuestas a estar dirigidas por la tecnología y la economía. Pero ¿qué ocurre si desafiamos estas suposiciones y abrimos los límites de la política a los no humanos? ¿Qué significa esto para la política de la IA y la robótica, y cómo puede la filosofía política y la teoría relacionada ayudarnos a conceptualizar esto?

Este capítulo aborda estas cuestiones considerando en primer lugar la teoría política que ya ha superado los límites humano/no humano: la teoría relacionada con los animales y el entorno natural. Presto especial atención a los argumentos relativamente recientes sobre el estatus político de los animales, en particular aquellos basados en los enfoques relacionales y poshumanistas. A continuación, el capítulo discute qué implica este cambio

en la teoría política para la política de la IA y de la robótica. Plantea dos tipos de preguntas. En primer lugar, considera el impacto de la IA sobre los no humanos (por ejemplo, los animales no humanos) y el entorno. ¿Debería el pensamiento político sobre la IA y los robots considerar las consecuencias para los no humanos, como los animales (no humanos) y los ecosistemas, dado que estos últimos pueden tener un estatus político? Si es así, ¿qué conceptos se pueden usar para justificarlo? Segundo, ¿pueden los sistemas de IA y los propios robots disponer de un estatus político? ¿Pueden llegar a ser, por ejemplo, un tipo de ciudadanos? ¿Cómo tratan esta cuestión el transhumanismo y el poshumanismo? Por ejemplo, ¿podría y debería la IA tomar el control político? ¿y cómo podemos volver a imaginar la política y lo social de una forma que incluya a los no humanos?

NO SOLO LOS HUMANOS CUENTAN POLÍTICAMENTE: EL ESTATUS POLÍTICO DE LOS ANIMALES (NO HUMANOS) Y DE LA NATURALEZA

En la ética animal y en la filosofía medioambiental encontramos propuestas para extender los límites de la consideración moral y política de los animales no humanos y del medioambiente. Dado el tema de este libro, me centraré en la consideración *política* y, además, limitaré el alcance de mi discusión a algunos de los principales argumentos sobre el estatus político de los *animales*, aunque mencionaré también la política medioambiental y la política del cambio climático.

Una obra bien conocida en esta área es *Liberación animal* de Singer (2009), que ofrece, como es muy conocido, un argumento utilitarista en favor de la liberación de los animales que usamos y matamos para la producción de comida, ropa y otras finalidades: deberíamos evaluar las consecuencias de estas prácticas para los animales; si les causamos sufrimiento, deberíamos reducirlo y, si fuera necesario, abolir por completo estas prácticas. Pero, a pesar de que Singer aborda este tema desde la ética, el libro se puede leer fácilmente como una obra filosófico-política que apela a principios políticos clave. Para comenzar, busca la liberación y se ha usado como justificación para liberar a los animales mediante el movimiento de liberación animal. El prefacio a la primera edición del libro

publicado en 1975 comienza con las palabras siguientes: «Este libro trata sobre la tiranía del ser humano sobre los animales no humanos» (8). Singer formula a continuación su ética usando un término político: tiranía, el opuesto a la libertad política. Pero no solo para apelar a la libertad política. También habla sobre la consideración de los intereses de los animales, extendiendo los principios de igualdad de los humanos a los animales no humanos, y con intención de acabar con el prejuicio y la discriminación, que tienen una larga historia. Su argumento central con respecto al especismo radica en una acusación de un tipo particular de sesgo y discriminación. El objetivo de liberar a los animales del abuso y la muerte en todos los ámbitos está justificado por el argumento de que, al igual que los humanos, muchos animales sienten el dolor y, por lo tanto, tratarlos de ese modo es «especismo»: la discriminación injusta de los animales no humanos basada en el simple hecho de que pertenecen a una especie diferente. Es «un prejuicio o actitud de sesgo en favor de los intereses de los miembros de la propia especie y contra los miembros de otras especies» (6). Y, debido a que la mayoría de nosotros (como carnívoros) somos los opresores y los que practicamos la discriminación, el cambio es difícil. Singer compara el movimiento de liberación animal con otros movimientos políticos importantes: el movimiento en favor de los derechos civiles, en los que la gente de color solo consiguió derechos porque los exigió (mientras que los animales no pueden hablar por sí mismos); la lucha por la abolición de la esclavitud, y el movimiento feminista con sus protestas contra el sexismo. Todo esto apoya su demanda de que son necesarias la protesta y la lucha. De este modo, el libro de Singer trata tanto de ética como de filosofía política. Su ética utilitarista, que es normalmente sobre lo que se centran los filósofos cuando responden a esta obra, se encuentra, de hecho, ligada a un conjunto de conceptos político-filosóficos que prácticamente abarcan el abanico completo que he cubierto en los capítulos anteriores. Nótese también que Singer adopta una posición universalista en su ética y en su filosofía política: no usa la identidad como un criterio, sino que argumenta sobre la base de una capacidad universal para el sufrimiento, independientemente de la especie.

Otros argumentos para otorgar consideraciones políticas a los animales, que también se encuentran dentro de la tradición liberal universalista, se basan en el principio político de la justicia. Por ejemplo, se puede utilizar un argumento contractual sobre la justicia para los animales. Una teoría influyente en esta área es la teoría de la justicia de Rawls. Rawls excluye a los animales de su teoría de la justicia (Garner, 2003), y este tipo de contractualismo normalmente descansa sobre la importancia de la racionalidad humana. Este es el motivo por el que Nussbaum (2006) escoge basarse en el concepto de las capacidades cuando se trata de otorgar estatus moral y político a los animales en términos de justicia: como los humanos, los animales también están dotados de capacidades para operar y florecer. Puede que no posean la racionalidad humana, pero poseen las capacidades específicas propias de sus especies y deberíamos respetar su propia dignidad. Nussbaum considera lo que implica su lista de capacidades/facultades para los animales. Por ejemplo, los animales tienen derecho a tener una vida sana, lo que implica que necesitamos leyes que prohíban el trato cruel (394). Esta es, por sí misma, una aplicación interesante de la teoría política: inicialmente destinada a los humanos, es ahora aplicada a los animales (nótese también que existe una relación interesante entre el enfoque de las capacidades de Nussbaum y la antigua ética y política de la virtud, centrada en el florecimiento de lo humano. En cualquier caso, no continuaré esta discusión aquí).

Sin embargo, algunos autores han propuesto que la justicia para los animales podría defenderse de acuerdo con una base contractual. Se podría modificar la posición original rawlsiana de tal forma que el velo de ignorancia incluya la ignorancia sobre si acabaremos como un humano o un animal. Por ejemplo, en la versión de Rowlands (2009), la racionalidad se encuentra oculta tras el velo de ignorancia, ya que es una ventaja natural innmerecida. Otra ruta sería subrayar la cooperación entre humanos y animales. He argumentado a favor de que los animales podrían incorporarse a la esfera de la justicia distributiva si son parte de un plan cooperativo (Coeckelbergh, 2009b): la idea es que los humanos y los no humanos son interdependientes de diferentes formas y que algunas veces cooperan. Si este es el caso, y cuando lo sea (por ejemplo, para los animales domésticos),

entonces estos animales deberían ser considerados como parte de la esfera de la justicia. Desplazando también el debate ético hacia el ámbito de la teoría política, Donaldson y Kymlicka en *Zoopolis* (2011) han argumentado de una forma similar que tenemos obligaciones relacionales hacia los animales y que deberíamos incluirlos en el proyecto cooperativo de la ciudadanía compartida. Los dos autores admiten que los animales pueden tener una capacidad reducida para la actuación política, pero pueden, no obstante, ser considerados como ciudadanos.

Esto es también lo que he denominado perspectiva «relacional» (Coeckelbergh, 2012): en lugar de propiedades o capacidades intrínsecas, Donaldson y Kymlicka subrayan las relaciones humano-animales. Estas relaciones nos prescriben obligaciones, por ejemplo el deber de cuidar de los animales dependientes de nosotros. Esto no implica que todos los animales tengan el mismo tipo de ciudadanía. Como en las sociedades humanas, argumentan los autores, algunos animales deberían ser miembros de pleno derecho de nuestras comunidades políticas (por ejemplo, los animales domesticados), mientras que otros tendrán sus propias comunidades (por ejemplo, los animales salvajes). Sin embargo, Garner (2012) ha criticado estos argumentos, ya que, según él, los autores que defienden que la posición original debería redefinirse se basan realmente en un principio que se encuentra fuera del contrato, y la confianza en la cooperación social funciona para los animales domesticados, pero no para los demás. Se podría replicar que (1) no está claro por qué estos contractualistas deberían atenerse a un estándar más elevado que el de Rawls, que también se refiere a juicios normativos preexistentes (en su caso, por ejemplo, la opinión de que la justicia solo se aplica a las personas), y (2) que los argumentos contractualistas están de hecho limitados cuando se trata de otorgar consideración *política* a los animales, pero que, por suerte, existen otros argumentos morales que disponen de un alcance de aplicación más amplio que pueden justificar protecciones más amplias, pero que no se centran en derechos y deberes *políticos* en sentido estricto. Podríamos otorgar consideración moral a muchos animales sobre la base de todo tipo de argumentos (por ejemplo, argumentos basados en la capacidad de sentir), pero no todos estos animales podrían tener un estatus

que permita considerarlos como beneficiarios de la justicia *política*. Por ejemplo, se podría defender que *podemos* tener deberes morales con, pongamos, un animal salvaje que está sufriendo en el bosque (por ejemplo, sobre la base de un deber de aliviar el sufrimiento al estilo de Singer y de la compasión), pero esta no es una obligación *política*, ya que el animal no es parte de nuestra comunidad política.

En respuesta a esta restricción, se podría exigir que la comunidad política tuviera un alcance más amplio que incluyera a todos los animales, a pesar de que sería difícil decir lo que esto implica. Por ejemplo, si otorgamos estatus político a los animales salvajes, ¿cuáles son nuestras obligaciones con ellos y en qué sentido, exactamente, son parte de nuestras comunidades políticas? Se podría ir más allá de los animales y discutir si, pongamos por caso, los ríos o los ecosistemas también deberían disponer de un estatus político, además de ciertos animales específicos. Por ejemplo, en 2018, la Corte Suprema de Colombia otorgó la condición de persona al bosque del Amazonas, y en Nueva Zelanda el río y el ecosistema Whanganui tienen un estatus legal. Esto último se puede justificar basándose en un argumento moral relacionado con su valor intrínseco, pero también se puede apoyar por medio de un argumento relacional, político, sobre la relación de interdependencia entre el río (por ejemplo, sus características espirituales) y el pueblo indígena iwi (y, por supuesto, ambos argumentos podrían estar conectados). Se puede mantener un debate similar sobre el valor de la Tierra o del planeta en su conjunto. Por ejemplo, en 2008, Ecuador decidió proteger los derechos de la Pachamama (la naturaleza) en su constitución, y en 2010, Bolivia (influenciada por las cosmovisiones andinas que consideran a la Tierra como un ser vivo) aprobó la Ley de los Derechos de la Madre Tierra, que define a la Madre Tierra como «un sujeto colectivo de interés público» y enumera diferentes derechos que posee esta entidad, incluyendo la vida y la diversidad de la vida (Vidal, 2011). Se podría interpretar esto como una muestra de respeto hacia el valor intrínseco de la Madre Tierra, o como una muestra de respeto de los derechos políticos del pueblo indígena, o ambas cosas.

En la ética medioambiental, existe una antigua discusión sobre el antropocentrismo (por ejemplo, Callicott, 1989; Curry, 2011; Næss, 1989;

Rolston, 1988) y el valor intrínseco (Rønnow-Rasmussen y Zimmerman, 2005). Por ejemplo, mientras que Regan (1983) restringe el valor intrínseco a los animales superiores, Callicott (1989), Rolston (1988) y Leopold (1949) defienden una ética que atribuye valor intrínseco a las especies, hábitats, ecosistemas y (en el caso de Rolston) a la biosfera. Este punto de vista va más allá de la filosofía moral antropocéntrica (que solo reconoce el valor intrínseco de los humanos) y se fundamenta en una comprensión ecológica de la naturaleza (McDonald, 2003). Mientras que estas discusiones se plantean en términos de ética, podrían expandirse hasta incluir las consideraciones políticas basadas en el valor intrínseco y usarse para justificar los derechos de las entidades naturales más amplias antes mencionadas.

La teoría poshumanista también ofrece una base para un enfoque no antropocéntrico y podría interpretarse que presta apoyo a la consideración política de animales no humanos. El *poshumanismo*, que significa tanto «después» como «más allá» del humanismo, se refiere a un cierto número de diferentes direcciones teóricas que son crítica del humanismo y la sociedad y cultura contemporáneas. El poshumanismo debe distinguirse del *transhumanismo*, que trata de la mejora de las capacidades humanas y que (al menos en una variante) considera la IA como reemplazo de los humanos o que toma el control del poder de los humanos (ver más adelante en este capítulo). En su forma filosófica, el poshumanismo deconstruye «lo humano» y las perspectivas jerárquicas y dualistas (Ferrando, 2019), y, por tanto, está en contra del antropocentrismo. Dado que esto cuestiona la posición privilegiada del ser humano en la tradición filosófica occidental y llama la atención sobre los no humanos y la hibridación, es posantropocéntrico, a pesar de que, ciertamente, no se puede reducir a esa posición (Braidotti, 2016, 14). Por ejemplo, en algunas versiones se hace un fuerte énfasis en la discriminación y la injusticia estructurales. Considérese el feminismo poshumanista y la teoría poscolonial. Se podría también reconocer que existe cierta variedad en la tradición occidental con respecto al antropocentrismo (por ejemplo, compárese a Kant y Hegel con Aristóteles y Marx): existen diferentes grados de antropocentrismo (Roden, 2015, 11-12). El posantropocentrismo y el antidualismo poshumanistas

están ligados: además de dualismos tales como sujeto/objeto y hombre/mujer, busca superar los dualismos humano/no humano, humano/animal, animado/inanimado, y así sucesivamente. Por lo tanto, también intenta atenuar el miedo de Occidente a la tecnología: en lugar de considerar las tecnologías como instrumentales o amenazantes, subraya lo que Derrida (1976; 1981) y Stiegler (1998), entre otros, denominan «tecnicidad originaria» de los humanos (ver también Bradley, 2011; MacKenzie, 2002, 3) y crea imaginarios de una vida conjunta con otros que son máquinas. El poshumanismo propone ontologías más abiertas que incluyan a la tecnología. De este modo, la IA deja de ser considerada como una amenaza a la autonomía humana: tras su trabajo de deconstrucción, ya no existe «lo humano» ni su autonomía no relacional, que podrían resultar amenazadas. El sujeto no tiene nunca el control total y es siempre dependiente de otros. Congruente con las explicaciones feministas de la autonomía relacional, tanto lo humano como el sujeto se consideran profundamente relacionales. Como ha defendido Rouvroy (2013) (siguiendo a Butler y a Althusser), no existe tal cosa como la autonomía total. Además, el poshumanismo es un proyecto filosófico, pero también político. Junto con los enfoques poscoloniales, foucaultianos y feministas (entre otros), por ejemplo en la obra de Haraway, Braidotti y Hayle, incluye las críticas a la violencia y a la totalización infligida a los animales no humanos por aquellos que mantienen una política antropocéntrica y cosmovisión excepcionalista de lo humano (Asdal, Druglitrø y Hinchliffe, 2017). Reconoce, además, que nosotros (humanos y no humanos) somos todos interdependientes y que todos dependemos de la Tierra (Braidotti, 2020, 27) (obsérvese que en la filosofía de la tecnología existen otros marcos éticos no antropocéntricos que provienen de diferentes direcciones teóricas, por ejemplo la ética de la información de Floridi [2013]).

Desarrollaré este conjunto de teorías poshumanistas con la pretensión de, primero, subrayar su actitud diferente frente a los animales y el entorno natural. Una figura fundamental del poshumanismo es Haraway. Después de su «análisis político-ficticio (político-científico)» en «El manifiesto cibernético» (2000), que ya cruzaba la frontera entre lo natural y lo artificial con la figura del cibernético («criaturas que son simultáneamente animal y

máquina»), Haraway defiende una política comprometida con el florecimiento de los animales de al menos dos formas. Primero, afirma que existen «especies de compañía» (Haraway, 2003), como los perros, y que nuestras relaciones, vida conjunta y, en última instancia, la coevolución con esos otros no humanos importantes conduce a la co-constitución de las de las identidades de ambos. Sobre la base de esta opinión, se podría defender que se otorgase al menos estatus moral y político a estos animales que cuentan como especies de compañía. Segundo, Haraway continúa abriendo las fronteras de lo político a los animales con sus conceptos de política *creando parientes (making kin)* y *multiespecies*. Respondiendo a la cuestión sobre el antropoceno, que subraya la agencia de los humanos a la hora de moldear la Tierra, Haraway defiende que no solo los humanos la han transformado, sino que hay otros «terraformadores», como las bacterias, y que existen muchas interacciones entre especies biológicas y tecnologías. Haraway (2015) piensa que la política debería promover el «florecimiento de ricos ensamblajes multiespecie que incluyan a personas», pero también a los «más-que-humanos» y a «otros-que-no-sean-humanos» (160). Basándose en este punto de vista «compuesto» (*composist*) (161), el cuerpo político se expande a todo tipo de entidades con las que podemos crear lazos de parentesco y ante quienes tenemos que responder. Como señala Haraway en *Seguir con el problema* (2016): somos responsables de «conformar las condiciones para el florecimiento multiespecie» (29), y esta respuesta crea vínculos que a su vez crean nuevos tipos de parentesco. Sin embargo, en una nota a pie de página advierte sobre la generalización y subraya el respeto por la diversidad y las situaciones históricas, creando un vínculo directo con la política humana (en particular, con la política estadounidense):

La creación de parentesco debe hacerse con respeto por los parentescos históricamente situados, parentescos diversos que no deberían ni generalizarse ni ser apropiados por el interés de una humanidad común, multiespecie colectiva o categoría similar demasiado rápida. [...] Es instructivo el lamentable espectáculo de muchos liberales blancos de los Estados Unidos que, ante el levantamiento de los afroamericanos y sus aliados organizándose contra los asesinatos policiales de personas negras y otros escándalos, oponen resistencia a #BlackLivesMatter insistiendo en que #AllLivesmatter es algo instructivo. Crear alianzas requiere reconocer especificidades, prioridades y urgencias. [...] Pretender crear un parentesco

mientras no se considera ni al colonialismo pasado ni presente ni otras políticas de exterminio y/o asimilación augura muchas «familias» disfuncionales, por decir algo (207).

Otro poshumanista, Wolfe, explora las implicaciones del concepto de biopoder y biopolítica de Foucault para las «relaciones transespecie» (Wolfe, 2010, 126). En *Before the Law: Humans and Other Animals in Biopolitical Frame* (2013), se pregunta por la exclusión de los animales en la tradición occidental desde Aristóteles hasta Heidegger. Critica el supuesto de Heidegger de que la humanidad y la animalidad son «zonas ontológicamente opuestas» (5-6), y usa el concepto de biopolítica, inspirado por Foucault (Wolfe, 2017), para defender que todos somos animales ante la ley:

La cuestión crucial de la biopolítica ya no es «humano» contra «animal»; esta cuestión se ha expandido recientemente hasta abarcar la comunidad de los vivos y las preocupaciones que todos nosotros deberíamos tener sobre dónde encajan la violencia y la protección inmunitaria, porque todos somos, después de todo, potencialmente animales frente a la ley (Wolfe, 2013, 104-105).

Massumi (2014) también se pregunta por la exclusión de los animales de lo político. Objeta la dicotomía humano/animal en el humanismo y la metafísica occidental y coloca a los humanos en el «*continuum* animal» (3). Cuestionándose la imagen que tenemos de nosotros mismos como seres que están separados de los demás animales, las prácticas de exclusión y mantenimiento de la distancia (por ejemplo, en el zoo, el laboratorio o en frente de una pantalla), y el pensamiento tipológico mismo con sus separaciones en categorías, propone una «política integralmente animal» en lugar de «una política humana de lo animal» (2). El foco está puesto en el papel del animal y en «convertirse-en-animal» (56-57): un término que está influenciado por Deleuze y Guattari (1987) y la filosofía del proceso, y que implica rechazar el pensamiento jerárquico y rígido sobre los humanos y los animales y problematizar la opresión de los animales no humanos.

De forma similar, Braidotti (2017) defiende repensar la subjetividad «como un ensamblaje colectivo que abarque a los actores humanos y a los no humanos, a la mediación tecnológica, a los animales, las plantas y al planeta en su conjunto» (9). Y añade una opinión específica normativa y política: necesitamos trabajar hacia una relación más igualitaria con los

otros no humanos (10) y rechazar «las configuraciones dominantes del ser humano como rey de la creación» (15). Braidotti propone un igualitarismo zoocéntrico, donde «zoo» se refiere a la fuerza vital de la vida no humana (16). Influenciada por Deleuze y Spinoza, defiende un enfoque monista y subraya el reconocimiento compasivo de la interdependencia con los otros no antropomórficos (22). Cudworth y Hobden (2018) también consideran el poshumanismo como un proyecto emancipatorio: buscan eliminar la centralidad del ser humano, pero sin perder la posibilidad de un compromiso crítico con las crisis de nuestro tiempo: los desafíos ecológicos y la desigualdad mundial (13). Su poshumanismo crítico explora alternativas al neoliberalismo (16-7) y es «tierrista» en el sentido de que es «una política para todo lo que vive» (136). Estamos integrados en un paisaje relacional y compartimos vulnerabilidades con otras criaturas y seres vivos. Influenciado por Haraway, Cudworth y Hobden escriben sobre la precariedad de los «bichos» (*critters*) encarnados vulnerables. (Para trabajos sobre la vulnerabilidad en los tiempos digitales desde un punto de vista más existencial, ver también Coeckelbergh, 2013 y Lagerkvist, 2019). Estamos llamados a imaginar un futuro más inclusivo más allá del neoliberalismo, más allá del antropoceno y del capitaloceno y (siguiendo a Latour) más allá de la modernidad (Cudworth y Hobden, 2018, 137).

Latour (1993; 2004) es conocido por su visión no moderna de la ciencia y de la sociedad, que cuestiona las distinciones humano/cosas y naturaleza/cultura cuando se trata de teorizar sobre lo social y lo político. De acuerdo con Latour, (el debate sobre) el calentamiento global es un híbrido, una mezcla entre política, ciencia, tecnología y naturaleza. Defiende una ecología política sin la idea de naturaleza. Inspirado por Latour e Ingold, también he cuestionado en varias publicaciones el uso del término «naturaleza» (por ejemplo, Coeckelbergh, 2015b, 2017). Además, como ha señalado Alaimo (2016), el término «naturaleza» «se ha utilizado durante mucho tiempo para apoyar el racismo, el sexismo, el colonialismo, la homofobia y el esencialismo» (11) y se encuentra, por tanto, lejos de ser políticamente neutral. Así, los poshumanistas han redibujado radicalmente las líneas de la comunidad política: no solo los humanos, sino también los no humanos (pueden) formar parte de ella. Esto no conduce necesariamente

a la ausencia de líneas divisorias o a la ausencia de exclusión, pero ya no acepta el dogma de que existe un profundo abismo entre humanos y no humanos cuando se trata de política. Los teóricos poshumanistas también cuestionan la distinción misma entre ética y política. De acuerdo con Alaimo,

incluso las prácticas éticas más pequeñas, más personales, de la esfera doméstica se encuentran estrechamente unidas a una gran cantidad de difíciles situaciones políticas y económicas, como el capitalismo global, la injusticia laboral y de clase, la injusticia climática, el neoliberalismo, el neocolonialismo, la agricultura industrial, la agricultura industrializada, la contaminación, el cambio climático y la extinción (10-11).

En cuanto tales, los poshumanistas defienden un enfoque menos independiente y están de acuerdo con los activistas del ecologismo que defienden que deberíamos cambiar nuestras vidas en lugar de hacer afirmaciones distantes sobre, por ejemplo, la «naturaleza» (muchos ecologistas continúan, sin embargo, refiriéndose a la «naturaleza», lo cual es una de las razones por las que no hay una superposición completa entre el ecologismo y el poshumanismo).

Algunos teóricos poshumanistas hacen uso de Marx. Por ejemplo, Atanasoski y Vora (2019) creen que Marx aún puede ser «aplicado a las fantasías tecnoutópicas de un mundo poscapitalista y poshumanista» (96) y lo combinan con un enfoque centrado en la raza, el colonialismo y el patriarcado. Un punto de vista interesante, que combina el marxismo con un enfoque ecológico, es el análisis de Moore del capitalismo-en-la-naturaleza, que cuestiona el binomio naturaleza/sociedad. Defiende que considerar la naturaleza como algo externo es una condición de la acumulación del capital. En cambio, deberíamos considerar el capitalismo como una forma de organizar la naturaleza (Moore, 2015). Critica a aquellos que hablan de cíborgs, ensamblajes, redes e híbridos por no escapar del pensamiento binario cartesiano.

También está relacionado con el poshumanismo el interesante concepto de política medioambiental de la justicia multiespecie, que cuestiona la narrativa antropocéntrica de la justicia: desafía el excepcionalismo humano y la idea de que los humanos están separados y son separables del resto de especies (Celermajer, *et al.*, 2021, 120). Tschakert (2020) subraya la

dimensión no humana del cambio climático y argumenta que la emergencia climática exige que los principios y prácticas de la justicia se revisen para abarcar tanto al mundo humano como al natural. Muestra las interconexiones entre la justicia climática y la de multiespecie mediante la exploración de encuentros, a través de los cuales se reconoce a «otros» no humanos con quienes se entretajan nuestras vidas (3). En la teoría legal existe también más reflexión sobre quién o qué pertenece a las «comunidades de justicia» (Ott, 2020, 94), por ejemplo sobre el estatus legal de los no humanos del antropoceno en relación con las injusticias dentro del sistema terrestre. Gellers (2020) propone ampliar las comunidades de justicia para incluir como sujetos legales tanto a los no humanos naturales como a los artificiales.

Ahora bien, si (¿algunos?) los animales y el entorno natural tienen estatus moral y político y si la comunidad política se abre para los no humanos, entonces, ¿qué implica esto para la política de la IA y la robótica?

IMPLICACIONES PARA LA POLÍTICA DE LA IA Y LA ROBÓTICA

Si dejamos atrás el antropocentrismo y expandimos la política a los animales y a la naturaleza no humana, se podrían desarrollar al menos dos tipos de posiciones con respecto a las potenciales implicaciones para la IA. Primero, se podría afirmar que su uso y desarrollo debería tener en cuenta a los animales, al medioambiente, etc., tener un estatus moral y político y, por tanto, evitar causar daño a estas entidades naturales y, si es posible, y preferentemente, contribuir de forma positiva a las prácticas más amigables medioambientalmente hablando y a solucionar problemas como el del cambio climático. Se podría promover una IA medioambiental y climatológicamente amistosa. En segundo lugar, se podría afirmar que la propia IA tiene un estatus político y que, por tanto, deberíamos hablar sobre la libertad, la justicia, la democracia, los derechos políticos, y así sucesivamente, *para la IA*, o, al menos, para ciertos tipos de IA (signifique esto lo que signifique). Presentaré estas posiciones y señalaré diferentes recursos teóricos disponibles para su posterior desarrollo.

EL SIGNIFICADO POLÍTICO DEL IMPACTO DE LA IA SOBRE LOS NO HUMANOS Y LOS ENTORNOS NATURALES

La primera posición es agnóstica sobre el estatus político de la propia IA (esto es, si la IA contaría ella misma como un no humano que necesita ser incluido en la comunidad política), pero afirma que, dado el valor político y los intereses de las entidades naturales, la política de la IA no puede ya ser concebida de una forma antropocéntrica. De acuerdo con esta posición, la IA debería no ser meramente antropocéntrica (esto es, estar orientada hacia el valor e intereses de los humanos), no digamos ya orientada hacia el capital, sino que debería también estar dirigida por el valor y los intereses de las entidades naturales como los animales, los ecosistemas y la Tierra. La cuestión crucial aquí no es que los animales puedan ser estudiados (por ejemplo, algunos investigadores de la IA se inspiran en la biología para diseñar los robots o adoptar sus métodos de organización social [Parikka, 2010]), sino que los animales tienen también valor político e intereses propios que deberían ser respetados. Cuando se desarrolla y usa la IA, se deberían tener en cuenta las consecuencias de la tecnología para los animales, el entorno natural y el clima.

Estas consecuencias no son necesariamente malas. La IA también ayuda a combatir el cambio climático y otros problemas medioambientales, por ejemplo cuando el uso del aprendizaje automático para analizar datos y ejecutar simulaciones mejora nuestro conocimiento científico del cambio climático y de los problemas medioambientales, o cuando ayuda a rastrear una extracción ilegal de recursos naturales. Y quizás la IA también pueda ayudarnos en relación con los animales, por ejemplo ayudándonos en la gestión y conservación de hábitats. Pero, al mismo tiempo, la tecnología también puede contribuir a que se den estos problemas. A veces, este aspecto es muy claro y visible. Por ejemplo, los asistentes personales que operan en el hogar con IA pueden golpear a las mascotas o confundirlas con su discurso. La IA usada para organizar la producción agrícola y cárnica puede conducir sistemáticamente a dañar a los animales, y la IA usada en la producción industrial puede tener efectos nocivos sobre el clima y el medioambiente. Sin embargo, incluso si todo esto es en principio visible

para el ojo del usuario u observador, la política antropocéntrica tiene un punto ciego para tales casos, ya que está centrada en la política para los humanos y, por tanto, también en la política de la IA para los humanos. Abrir la política a los no humanos nos ayuda a revelar, imaginar y discutir estos problemas.

Sin embargo, el uso de la IA tiene a menudo consecuencias que son más remotas y menos visibles desde el punto en el que la usamos. Hay aquí un tema importante ya mencionado anteriormente, pero que merece más atención: el del impacto de la IA sobre el medioambiente y el cambio climático en términos de uso de energía y de recursos. La computación necesaria para algunos tipos de aprendizaje automático ha aumentado dramáticamente, y esto consume mucha energía, que a menudo no se extrae de fuentes renovables. Por ejemplo, el entrenamiento de redes neuronales para el procesamiento del lenguaje natural (PLN) necesita recursos considerables debido a su uso de la electricidad y (por tanto) genera una considerable cantidad de huella de carbono. Entrenar un único modelo de PLN puede conllevar emisiones de dióxido de carbono equivalentes a más de cinco veces la cantidad producida por un coche medio durante toda su vida útil (Strubell, Ganesh y McCallum, 2019). Actualmente, los investigadores están intentando encontrar formas para lidiar con este problema, por ejemplo entrenando con menos datos o, incluso, reduciendo el número de bits necesarios para representar los datos (Sun, *et al.*, 2020). Las principales empresas de IT como Apple y Google han firmado compromisos con las energías renovables. Sin embargo, la mayoría de las empresas tecnológicas aún dependen de combustibles fósiles y, como sector en conjunto, tienen una importante huella de carbono global. De acuerdo con un informe de Greenpeace de 2017, la huella de carbono del sector se estimaba que era del 7 por 100 de la electricidad global y se esperaba que aumentase (Cook, *et al.*, 2017; ver también AI Institute, 2019). Usar dispositivos de transmisión (*streaming*) en combinación con sistemas de recomendación de IA es parte del problema, ya que necesita de más datos. La producción de dispositivos electrónicos que trabajan con IA también requiere de la extracción de materias primas, lo que conlleva consecuencias sociales y medioambientales. Al igual que las formas históricas del

colonialismo, el «colonialismo de datos» (Couldry y Mejias, 2019, xix) va de la mano de la explotación de los recursos naturales: la actual explotación de los seres humanos no está ocurriendo «en lugar de los recursos naturales» (xix), sino además de ellos, y basándose en la extracción de los recursos naturales. Por lo tanto, el uso de la IA y de otras tecnologías digitales no conduce necesariamente a una desmaterialización de la economía, como algunos esperan. También conduce a un mayor consumo y, por tanto, a una mayor «presión ecológica» (Dauvergne, 2020, 257). De forma similar, el capitalismo de la vigilancia no trata solo sobre la destrucción de la dignidad humana, sino que tiene también una cara medioambiental. El hecho de que estas consecuencias medioambientales y climáticas no sean visibles en el momento de uso, sino que ocurren a menudo en lugares remotos del uso particular de la IA, no hace que sean menos importantes o problemáticos políticamente hablando. Además, a pesar de que, como hemos mencionado, la IA puede ayudar a combatir el cambio climático (Rolnick, *et al.*, 2019), esto no necesariamente compensa los problemas indicados.

Basándose en este análisis, se podría argumentar de forma normativa que el impacto de la IA en términos de huella de carbono y consecuencias medioambientales es políticamente problemático por dos razones. Primera, el cambio climático y el agotamiento de los recursos naturales conllevan consecuencias para los humanos y para las sociedades humanas, de las cuales se supone normalmente que tienen un estatus político y que son dependientes de los recursos naturales y de las condiciones climáticas. Segundo, basándose en los argumentos relacionados con los intereses políticos, el valor intrínseco, etc., de los *no humanos*, se podría añadir que esto también es problemático debido a su estatus político. Si unos animales específicos, los ecosistemas y la Tierra en su conjunto son afectados por la IA (bien directamente a través de, por ejemplo, la extracción de materias primas para dispositivos de IA o por medio de la gestión controlada por IA de animales criados para el consumo, o bien indirectamente a través de la destrucción de hábitats y el impacto multifacético del cambio climático debido a la producción de electricidad basada en carbón, necesaria para el funcionamiento de la IA), entonces este impacto es políticamente

problemático porque los *animales*, *entornos*, etc., y no solamente los humanos, cuentan y deberían contar políticamente.

Con todo, la idea de un desplazamiento hacia una política de la IA no antropocéntrica no debería considerarse como si se tratase meramente de un valor político o de un principio como «no humanos», «el clima» o «el medioambiente». Al contrario, resuena a través de todos los principios y discusiones que hemos examinado y construido en capítulos anteriores. Por ejemplo, si (algunos) animales cuentan políticamente, entonces deberíamos también considerar *su* libertad, hablar sobre justicia interespecies, preguntarnos si estos animales también podrían ser ciudadanos en una democracia, y así sucesivamente. Entonces, la política de la IA pasaría de ser antropocéntrica a estar centrada en los intereses y necesidades que compartimos con otros animales, y en los intereses y necesidades específicos que no compartimos con ellos, pero que continúan siendo políticamente importantes. Un giro no antropocéntrico hacia la política de la IA no significaría contemplar a los animales, al medioambiente y al clima como una consideración adicional en un marco ya existente, sino que constituiría un cambio fundamental sobre la idea misma de lo político, que se ha expandido hasta incluir a los no humanos y sus intereses.

Además, la geopolítica (definida aquí como la política de la Tierra o del planeta) necesita ser redefinida una vez que eliminamos a los humanos del centro. Desde la perspectiva de una visión no antropocéntrica y en línea con las ideas poshumanistas antes mencionadas, ya no es apropiado hablar del «antropoceno», ya que puede sugerir que los humanos están en el centro y son los que mandan, o que *deberían* estar en el centro y ser los que mandan. Al contrario, es importante subrayar que compartimos la Tierra y el planeta con muchos otros seres. Puede ser que los humanos hayamos asumido una capacidad muy importante de acción y que tratemos al planeta y a sus habitantes humanos y no humanos como su nave espacial, que puede y debe ser gestionada y quizás rediseñada con la ayuda de la IA y de otras tecnologías. Pero este es el *problema*, no la solución. Una política no antropocéntrica de la IA sería menos tecnocéntrica y problematizaría ese tipo de actitud y forma de hacer las cosas: implicaría cuestionar que la IA pueda ser una solución para todos nuestros problemas y, al menos,

considerar la idea de debilitar nuestro dominio sobre el planeta, en lugar de incrementar los medios de control de la IA y de la ciencia de datos. Una vez que ponemos en cuestión el tecnosolucionismo (Morozov, 2013) o lo que Broussard (2019) denomina «tecnochauvinismo» (la creencia en que la tecnología tiene siempre la solución), dejaremos de ver a la IA como la solución mágica a todos nuestros problemas y prestaremos más atención a los límites de lo que puede hacer: para nosotros y para los no humanos.

¿ESTATUS POLÍTICO PARA LA PROPIA IA?

La segunda posición no contempla a la IA como un simple instrumento al servicio de los humanos o, incluso, de los no humanos, sino que considera también la posibilidad de que algunos sistemas de IA obtengan un estatus político, por ejemplo en la forma de «máquinas», signifique esto lo que signifique. La idea es que la IA no es meramente, o no debería ser, una forma tecnológica de alcanzar un fin político, un instrumento para la política, sino que puede ser más que un instrumento que *puede tener un valor político e intereses propios*, y que, bajo ciertas condiciones, debería estar incluida en la comunidad política. Lo que aquí se afirma es que la política de la IA debería ser no solo política de humanos *para* no humanos (y humanos), teniendo el «para quién» como una pregunta abierta que también incluya potencialmente a los no humanos tecnológicos.

En la filosofía de la tecnología, este tipo de cuestión normalmente se formula como una cuestión *moral*: ¿pueden tener las máquinas estatus moral? Durante las últimas dos décadas, ha existido un intenso debate sobre este tema en la comunidad ética de la IA y la robótica (por ejemplo, Bostrom y Yudkowsky, 2014; Coeckelbergh, 2014; Danaher, 2020; Darling, 2016; Floridi y Sanders, 2004; Gunkel, 2018; Umbrello y Sorgner, 2019). Pero ¿qué es lo que sucede con el estatus *político* de estas entidades? ¿Qué es lo que sucede con sus derechos políticos, su libertad, su ciudadanía, su justicia con respecto a cómo nos relacionamos con ellas, etc.?

Hay aquí diferentes e interesantes rutas para explorar algunas afirmaciones y argumentos en esta dirección. Una primera forma es sacar provecho del debate sobre el estatus moral de las máquinas. Parte de la

discusión sobre el estatus político probablemente será una imagen especular de ese debate. Por ejemplo, aquellos que son reacios a otorgar estatus moral a la IA sobre la base de una falta de propiedades como la conciencia o la capacidad de sufrimiento probablemente usarán los mismos argumentos para negar el estatus político a la IA. Otros pueden argumentar que tales propiedades intrínsecas no son importantes para el estatus político y que lo que más bien cuenta es la dimensión relacional o comunal: si nos relacionamos con la IA de forma social y comunal, ¿no debería ser esto suficiente para otorgarles algún estatus político a estas entidades? Por ejemplo, Gunkel (2018) y yo mismo (Coeckelbergh, 2014) hemos cuestionado adscribir estatus moral sobre la base de propiedades intrínsecas y hemos sugerido que dónde trazar la línea es una cuestión política de (re)negociación, que está vinculada a la historia humana de exclusión y poder y que debería hacernos ser prudentes ahora con respecto a adscribir estatus moral a otras entidades. Lo mismo podría decirse, *a fortiori*, para el estatus político. ¿Es tiempo de renegociar la exclusión de las máquinas de la política? Una distinción que también es relevante para el estatus político es la que hay entre la agencia y la paciencia moral: la agencia moral trata sobre lo que la IA debería (o no) hacer hacia otros, mientras que la paciencia moral trata sobre lo que se le puede hacer a la IA. Se podría hacer una distinción similar entre la agencia y la paciencia política de la IA. Por ejemplo, podríamos preguntarnos si la IA tiene las propiedades necesarias para el juicio y la participación política (agencia), o qué es *debido a* la IA, políticamente, bajo ciertas condiciones (paciencia).

Se puede hacer una conexión interesante entre esta discusión sobre el estatus moral y la filosofía política (y legal) centrándose en el concepto de «derechos». El libro de Gunkel *Robot Rights* (2018), que discute sobre el estatus moral y legal de los robots para examinar su situación social, es un punto de partida excelente para este ejercicio. Además de ofrecer una útil clasificación de las diferentes posiciones sobre los derechos de los robots, Gunkel va más allá del antropocentrismo al señalar que los derechos no significan necesariamente derechos *humanos*, y que creemos erróneamente que sabemos de lo que estamos hablando cuando usamos la palabra «derechos». Existen diferentes tipos de derechos, como privilegios,

reclamaciones, poderes e inmunidades. Mientras que la mayoría de la gente rechazará otorgar derechos *humanos* a los robots o a la IA, se podría usar este análisis e intentar defender un tipo particular de derecho para las máquinas que sea menor que, o diferente de, los derechos humanos, y considerar su significado *político*. El reconocimiento de que los derechos no necesitan ser humanos, o estar al nivel de los de los humanos, abre algo de espacio para discutir esta cuestión. Sin embargo, un proyecto así no sería solo rechazado por aquellos que se oponen *a priori* a la idea de los derechos de los robots; puede encontrarse también con objeciones relacionales. En línea con el poshumanismo crítico y el posmodernismo, se podría defender (Coeckelbergh, 2014; Gunkel, 2018) que este tipo de razonamiento constituye él mismo una forma de filosofía totalizadora que coloca a la filosofía moral y política sobre un tipo de plataforma a partir de la cual se categorizan las entidades y se les otorgan derechos, morales y políticos. Nótese también que aplicar el derecho romano para otorgar a los robots un estatus político comparable a los esclavos, o de forma más general, modelando un marco de derechos políticos basándose en categorías de derecho romano de derechos y ciudadanía, sería igualmente problemático, si no más. También lo sería desde un punto de vista marxista, poscolonial y de la política identitaria (ver capítulo 3).

Otra ruta es tomar los argumentos relacionados con el estatus político de los animales y el medioambiente y tratar de aplicarlo a la IA, por ejemplo aquellos que ligan el estatus político con la cooperación o el valor intrínseco (véase la primera sección de este capítulo). Si la conciencia o la capacidad de sentir son vistas como una propiedad intrínseca que garantiza el estatus moral o político en el caso de los animales, entonces, si una entidad artificialmente inteligente fuese capaz de mostrar la misma propiedad, tendríamos que otorgarle el mismo estatus. Si la IA, como algunos animales, llegase a considerarse que tiene intereses o que actúa como un colaborador en un esquema cooperativo, entonces, de acuerdo con, por ejemplo, los argumentos de Donaldson y Kymlicka (2011), podría considerarse como candidata a tener estatus político, incluso si no es consciente o sentiente (pero nótese que existen debates considerables sobre qué cuenta como tener capacidad de sentir). Y, si ciertas entidades no

humanas como los ríos o las rocas pueden disponer de estatus político sobre la base de su valor intrínseco, sin tener que ser conscientes o sentientes, entonces, ¿podrían algunos tipos de IA o «ecologías» de IA recibir un estatus similar, especialmente si se consideraran como sagradas o especialmente valiosas para una comunidad en particular? Además, dadas las posibilidades de la biología sintética y la bioingeniería, los límites entre entidades vivas y no vivas son, cada vez, más difusos. Los laboratorios de los científicos producen híbridos entre organismos y entidades artificiales. En el futuro, puede que tengamos más «organismos sintéticos» y «máquinas vivas» (Deplazes y Huppenbauer, 2009). Por ejemplo, los bioingenieros intentan producir células programables; las células parecen vivas y, al mismo tiempo, son máquinas, ya que son un producto del diseño. Si, y hasta qué punto, la IA se mueve en la dirección de la «vida» (por ejemplo, se convierte en una «máquina viva»), entonces, si la propia vida es considerada como condición suficiente para el estatus político (por ejemplo, paciencia política), cumpliría, al menos hasta cierto punto, las mismas condiciones para el estatus político que (otros) seres vivos. Y en cualquier caso, ¿hasta qué punto es importante la categoría «vida»? Si las especies y entornos naturales particulares son valiosos por sí mismos y si eso permite asignarles derechos políticos como defienden algunos ecologistas, entonces, ¿deberían las «especies artificiales» y los «entornos artificiales» (Owe, Baum y Coeckelbergh por publicar) estar protegidos políticamente si cumplen funciones similares y se perciben como intrínsecamente valiosos? Los poshumanistas pueden también acoger todo tipo de híbridos en la comunidad política. Para ellos, las máquinas vivas y la vida artificial no son un problema: al contrario, aferrarse al antiguo dualismo vida/no vida es problemático, y desestabilizar tales dualismos y fronteras es parte de su proyecto filosófico y, ciertamente, *político*.

Como algunos argumentos sobre el estatus moral de las máquinas, ciertos argumentos sobre el estatus político pueden verse alentados por el reciente éxito en el que la IA ha superado a los humanos (por ejemplo, en el ajedrez y en el go) y se puede aceptar que en el futuro los humanos desarrollarán una inteligencia artificial general (IAG) o superinteligencia (SIA). Transhumanistas como Bostrom (2014), Kurzweil (2005) y Moravec

(1988) consideran seriamente esta posibilidad y discuten escenarios en los que aparecen inteligencias superiores a la humana, colonizan la Tierra y se expanden por el universo. En relación con tales escenarios, la cuestión política no se preocupa solamente del estatus político de estas entidades superinteligentes (como agentes o pacientes políticos), sino también del estatus político de los humanos: ¿se hará cargo una IA superinteligente del poder político de los humanos, y qué implicaciones tendrá esto para su estatus político? ¿Serán los humanos esclavos de sus amos artificiales, como son ahora algunos animales esclavos de los humanos? ¿Serán usados los humanos como material biológico o como fuentes de energía (como hacemos actualmente con algunas plantas o animales, ver también la película *Matrix*) o se les cargará en un dominio digital? ¿Existirán siquiera los humanos entonces? ¿Cuáles son los riesgos existenciales para la humanidad (Bostrom, 2014), usando aquí el término «existencial» en el sentido de que la humanidad podría ser destruida? ¿Podemos asignar metas a la IA que sean compatibles con las de los humanos o cambiará estas metas y buscará las suyas propias, en lugar de preocuparse de los intereses humanos? La primera opción es la favorecida por transhumanistas como Hughes (2004), que (en contraste con los poshumanistas) busca continuar y radicalizar no solo la Ilustración, sino también el humanismo y la democracia, incluso a costa de alterar radicalmente lo humano (Ferrando, 2019, 33). Otros creen que lo humano debe ser completamente superado y que es mejor que hagamos espacio para seres más inteligentes que nosotros. En cualquier caso, los transhumanistas están de acuerdo en que tecnologías como la IA, más bien que la cultura y educación humanas, crearán y deberían crear mejores humanos y más-que-humanos, y que esto conllevará consecuencias políticas para la humanidad en su conjunto en el sentido de que probablemente desplace el equilibrio de poder entre humanos (si es que queda alguno) y sus amos mejorados o superinteligentes. Puede que dejen de ser necesarias la democracia y otras antiguas instituciones, como ha sugerido Harari (2015). Mientras que la mayoría de los transhumanistas emplean menos tiempo en reflexionar sobre las consecuencias políticas más inmediatas para los humanos y las sociedades humanas concretas, también existen ramas del transhumanismo que son explícitamente políticas y se

preocupan del presente y que abarcan desde los libertarios (Zoltan Istvan, quien se presentó a presidente de los Estados Unidos en las elecciones de 2016) hasta los demócratas y los izquierdistas, que llaman la atención sobre los problemas sociales (James Hughes).

Menos futurista y ya actualmente relevante es el papel de la IA en la política y la gobernanza como agente político. Considérese de nuevo la cuestión del conocimiento experto y el liderazgo político: ¿qué tipo de juicio y capacidades se necesitan para la política, y cómo podemos equilibrar tecnocracia y democracia? Si los humanos ejercen su juicio sobre desafiantes cuestiones políticas, y si esta capacidad está relacionada con su autonomía y estatus normativo, ¿podría la IA desarrollar tal capacidad de juicio y, por tanto, adquirir agencia política? ¿Podría la IA adoptar la política de los humanos? ¿Tendría el tipo de conocimiento, conocimiento experto y capacidades necesarias para el liderazgo político? Y ¿sería compatible tal rol de la IA en política con la democracia? Ya discutimos alguna de estas cuestiones en el capítulo 4, pero ahora nuestra preocupación no es meramente el papel instrumental de la IA en relación con el conocimiento, la tecnocracia y la democracia, sino con si puede, y cómo, adquirir su propia agencia política.

Para responder a esta cuestión, podemos volver atrás a la discusión sobre el estatus moral y, en particular, la agencia moral de los no humanos (por ejemplo, de nuevo Floridi y Sanders), pero también podríamos desarrollar esta discusión considerando de nuevo el pensamiento poshumanista sobre los no humanos (Latour), o, como en el capítulo 4, refiriéndonos a la teoría política sobre el liderazgo y la ciudadanía, sobre si son necesarias las capacidades racionales, cuál es el papel de las emociones, qué es el conocimiento experto en política, y así sucesivamente. Considérese, por ejemplo, el aviso de Arendt de que el sentido común, la reflexión y el juicio son necesarios en política. Siguiendo mis sugerencias en el capítulo 2 y 4, se podrían plantear objeciones sobre si la IA podría adquirir estas capacidades.

Además de la pregunta sobre si «podría», está también, por supuesto, la pregunta de si «debería». Pocos autores son abiertamente entusiastas sobre la posibilidad de una tecnocracia de la IA. Incluso dentro de la literatura

transhumanista, la tecnocracia es a menudo descartada. De forma más destacada, Hughes ha defendido en *Citizen Cyborg* (2004) que la tecnología debería estar bajo nuestro control democrático. Además de la razón, de la ciencia y de la tecnología, necesitamos la democracia. Más que eso, Hughes insta a que el dominio tecnológico sobre la naturaleza «requiera de una democratización radical» (3). Defiende una forma democrática de transhumanismo y superinteligencia: daremos a luz otras formas de inteligencia, pero deberíamos ahora «acabar con la guerra, la desigualdad, la pobreza, las enfermedades y la muerte innecesaria» porque ello determinará la forma en que moldeemos el futuro transhumanista (xx). Además, la idea de que los humanos actuales diseñen cómo serán los seres humanos en el futuro plantea problemas en relación con la justicia transgeneracional que son relevantes aquí y ahora: ¿queremos tal responsabilidad? Sin embargo, muchos transhumanistas prefieren centrarse en el futuro distante, un futuro que no tenga lugar necesariamente en esta Tierra y con la humanidad tal como la conocemos. Actualmente, los millonarios tecnológicos como Elon Musk y Jeff Bezos parecen apoyar tales escenarios y hacen planes para colonizar el espacio.

Parte de la teoría poshumanista aborda la idea del estatus político de las máquinas y la de ampliar hacia ellas la comunidad política. En la sección previa ya señalé cómo Haraway defiende una política que traspase los límites organismo/máquina y los límites biológicos especies/tecnología, en contra de los órdenes y concepciones binarias hegemónicas. Esto es relevante no solo para el estatus político de los animales, sino también para el estatus político de las máquinas. La metáfora del cibernético está destinada a deconstruir el dualismo humano/máquina. Haraway presenta una «política feminista (y poshumanista) que abraza su entrelazamiento con las nuevas tecnologías» (Atanasoski y Vora, 2019, 81. Cursiva del traductor). Sin embargo, también aquí la cuestión no es solo que necesitemos reflexionar de forma diferente sobre las tecnologías: también necesitamos repensar a los humanos y la política. Gray (2000) adopta el imaginario del cibernético de Haraway en *Cíborg Citizen* (2000). El término se refiere a la observación de que los humanos como especie continúan transformándose tecnológicamente: en este sentido, somos cibernéticos. La cuestión es, entonces:

¿cómo sería una «sociedad cíborg» (2)? ¿Qué significaría la ciudadanía en «la edad de la reproducción electrónica» (21)? El autor afirma que las tecnologías son políticas y que el orden tecnológico debe ser más democrático (198). Dado que el conocimiento es poder, los ciudadanos cíborgs necesitarán información para gobernar. También necesitamos nuevas instituciones políticas cuando se están desarrollando nuevas tecnologías.

Como Haraway, Barad (2015) tiene interés en imaginarios políticos nuevos. Apoyándose en *Frankenstein* de Shelley y en las teorías *queer* y *trans*, argumenta que los monstruos pueden invitarnos a explorar nuevas formas de devenir y parentesco (140) e imagina una fusión con el otro no humano y poshumano. En *Meeting the Universe Halfway* (2007), cuestiona la división del mundo en categorías tales como lo social y lo natural. Usando la mecánica cuántica como fuente para sus metáforas argumenta que, por el contrario, deberíamos «teorizar lo social y lo natural conjuntamente» (24-25). Intenta hacerlo con su versión del realismo, un realismo agente, que también responde a un entendimiento *performativo* de las prácticas del discurso (ver de nuevo Butler). De acuerdo con Barad, hay formas de agencia humanas y no humanas. Conecta esto con una noción del poder inspirada por Foucault y Butler, pero (inspirada por Fernandes) habla también sobre las relaciones de producción, que están configuradas por una agencia maquina. Las máquinas y los humanos emergen a través de «entrelazamientos específicos de agencias» (239). Se constituyen unos a otros. En otras partes, Barad (2003) critica el representacionalismo y lo enlaza con el individualismo político. Su alternativa es la noción de *performatividad* poshumanista, que pone en entredicho categorías como las de humano/no humano (808) o social/material, y entiende el poder no solo como algo social, sino también como operativo en las materializaciones (810). Tanto los cuerpos humanos como los no humanos «llegan a ser materia» a través de la *performatividad* y las «intra-acciones agenciales» (823-824).

Basándonos en tales enfoques, muchos de los fenómenos discutidos en los capítulos previos pueden ser replanteados de tal manera que incluyan no solo a los humanos, sino también a los no humanos. La producción de

sesgos, la desigualdad y las formas totalitarias de control y vigilancia pueden ser teorizadas de modo que se basen no solo en los humanos, sino también en los no humanos, incluyendo a las tecnologías y a las instituciones. Por ejemplo, influenciados por Deleuze y Guattari, Haggerty y Ericson (2000) han defendido que, en lugar de utilizar la metáfora del panóptico para describir la vigilancia contemporánea, deberíamos usar el término «ensamblaje vigilante», ya que implica tanto a los humanos como a los no humanos.

Otro enfoque un tanto poshumanista y, ciertamente, deconstructivo del dualismo humano/no humano con respecto a la IA consiste en afirmar que cuestionar la distinción humano/IA no es cuestionar dos términos que son fijos: en el proceso, los dos también cambian. Cuando discutimos qué es lo que entendemos por IA, no discutimos solamente sobre una tecnología, sino que estamos usando y discutiendo una metáfora: al usar el término «inteligencia artificial» nos basamos en una comparación con la inteligencia humana. Ahora bien, las metáforas tienden a cambiar los dos términos que conectan. Inspirado por la concepción de Ricoeur sobre la metáfora y, en concreto, su término «asimilación predictiva», que plantea que la metáfora genera «nuevos lazos de semejanza que no existían antes de la unión metafórica» (Ricoeur, 1978, 148), Rhee (2018, 10-11) ha identificado un «colapso metafórico» entre los humanos y las máquinas con respecto a la IA: la antropomorfización de la IA crea un humano que no existía antes de la humanización. En otras palabras, la IA nos hace reflexionar de forma diferente no solo sobre las máquinas, sino también sobre los humanos. Esto también tiene implicaciones políticas. Siguiendo a Rhee, podemos decir que decidir cómo conformar las relaciones entre humanos y la IA es un acto de poder. La creación y el uso del término «IA» son, ellos mismos, actos políticos en este sentido. Las metáforas promueven también creencias específicas. Por ejemplo, Broussard (2019) ha defendido que el término «aprendizaje automático» sugiere que los ordenadores tienen agencia y que son entidades sentientes porque aprenden. Tal «confusión lingüística» (89) puede también ser vista como un ejercicio (*performativo*) del poder.

A veces el poshumanismo lleva a cabo una conexión explícita con la teoría política. Zolkos (2018) ve un giro poshumanista en la teoría política

que entraña una teorización no antropocéntrica de la política. Esto significa, entre otras cosas, que los organismos biológicos y las máquinas son considerados en términos de su agencia política. Por ejemplo, Lazzarato (2014) une lo simbólico (signos) y lo tecnológico (máquinas). Las máquinas son actores sociales, y las megamáquinas (un término prestado del filósofo de la tecnología Lewis Mumford) son ensamblajes que incluyen humanos, animales no humanos y objetos inanimados. Lazzarato argumenta que en el capitalismo tardío los humanos están sujetos al funcionamiento de las megamáquinas. La política, entonces, no es solamente una cuestión de humanos, sino que ocurre adicionalmente dentro de las (mega)máquinas, donde los humanos, pero también las máquinas, objetos y signos son agentes, y aquello a través de lo cual se produce la subjetividad. Como «sugieren, permiten, solicitan, impulsan, incitan y prohíben ciertas acciones, pensamientos y afectos» (30), las máquinas establecen relaciones de poder: el tipo de relaciones de poder sobre las que teorizó Foucault. En particular, Lazzarato contempla la *esclavitud* maquinal: se trata de «el modo en el que funcionan la ciencia, la economía, la comunicación de redes y el bienestar del Estado» (31). Critica el neoliberalismo y ve la posibilidad de un cambio político radical en el caso de sucesos graves e inesperados.

Sin embargo, ¿puede el transhumanismo e, incluso, el poshumanismo, apoyar de forma suficiente tal cambio y adoptar un papel crítico? Desde la perspectiva de una teoría crítica, se podría considerar toda esta discusión sobre el estatus moral y político de la IA como narrativas y *performances* de ciencia ficción que se arriesgan a apoyar el uso y desarrollo de la IA para generar beneficio para unos pocos mediante la explotación capitalista. En el capítulo previo sobre el poder puse el ejemplo del robot humanoide Sophia: sus *performances* y narrativas apelan a la idea de estatus político (ciudadanía), pero (o así podría argumentar un teórico crítico) aquí realmente de lo que se trata es de conseguir beneficios y acumular capital. Desde luego que deberíamos discutir el impacto de la IA sobre los seres humanos y, quizás, también sobre los no humanos, pero hablar sobre el estatus político de la IA (como agentes o pacientes) puede distraernos de pensar en las formas de explotación capitalista que son malas para la gente y el (resto del) planeta.

Finalmente, como nos recuerdan poshumanistas, ecologistas y feministas, hablar sobre el futuro de la «humanidad» y de nuestra relación con la «naturaleza» o, incluso sobre los «otros» no humanos y las máquinas (como sucede en la obra de algunos poshumanistas), puede distraernos de contemplar la política al nivel de la esfera doméstica y personal: la «pequeña» política que está relacionada con la «gran» política. Lo que hacemos con la IA y con nosotros mismos en la vida cotidiana también es político (y, como mostraron las discusiones sobre la igualdad y el poder en los capítulos previos, cómo definamos el «nosotros» es algo político). La política de la IA tiene un alcance muy profundo en cuanto a lo que tú y yo hacemos con la tecnología en nuestro hogar, en el lugar de trabajo, con los amigos, etc., que, a su vez, conforma la política. Quizás sea este el poder real de la IA: mediante lo que hacemos en nuestros teléfonos inteligentes y en otras pantallas en nuestra vida (mundos) cotidiana, otorgamos poder a la IA y a aquellos que la usan para la acumulación de capital, apoyamos estructuras sociales hegemónicas específicas, reforzamos divisiones binarias y negamos pluralidades. En este sentido, también, «los datos son el nuevo petróleo»: ni los datos ni el petróleo serían tan importantes si no los usáramos, si no fuéramos adictos a ellos. De nuevo, la IA y la política de la IA tratan sobre la gente: la gente que tiene interés en mantenernos adictos a ella. Al mismo tiempo, esto también abre posibilidades para la resistencia y el cambio.

CAPÍTULO 7

Conclusión: tecnologías políticas

QUÉ HE HECHO EN ESTE LIBRO Y QUÉ PODEMOS CONCLUIR

En respuesta a los problemas normativos planteados por la IA y algunas tecnologías relacionadas, este libro ha defendido que cuando se trata de usar recursos desde la filosofía práctica es útil recurrir no solo a la ética, sino también a la filosofía política. En cada capítulo he hecho sugerencias sobre cómo tender puentes entre la IA y la filosofía política, centrándome en principios y problemas políticos específicos y relacionándolos con la IA. Ha quedado claro que los problemas de los que nos preocupamos actualmente en las discusiones políticas y sociales, tales como la libertad, el racismo, la justicia, el poder y (las amenazas a) la democracia, adquieren una nueva urgencia y significado bajo la luz de los desarrollos tecnológicos como la IA y la robótica, y que la filosofía política puede ayudar a conceptualizar y discutir estos problemas y significados. El libro muestra cómo las teorías sobre la libertad, la justicia, la igualdad, la democracia, el poder y la política no antropocéntrica pueden ser usadas productivamente para reflexionar sobre la IA.

De forma más precisa, he llevado a cabo un ejercicio con dos vertientes. Por una parte, he mostrado cómo conceptos y teoría procedentes de la filosofía política y de la teoría social nos pueden ayudar a abordar, comprender y plantear los desafíos normativo-políticos planteados por la IA. Esto ha conducido a formular un bosquejo de *una filosofía política de la IA*, que es *una* especie de caja de herramientas conceptual que puede ayudarnos a reflexionar sobre la política de la IA. No pretende ser exclusivo: recibo con los brazos abiertos los esfuerzos llevados a cabo desde otras direcciones, y parte de los textos citados no provienen de la filosofía política, en sentido estricto. Además, el libro pretende ser una

introducción y deja sitio suficiente para trabajos posteriores en cada uno de estos temas. Sin embargo, *ofrece unos sustanciales cimientos para un marco evaluativo y normativo desde el que puede producirse la reflexión sobre los aspectos políticos de la IA* y que pueden ser útiles para aquellos que estén interesados en los aspectos normativos de la IA en la investigación, la educación superior, los negocios y la política. Espero que sus herramientas y discurso conceptuales no sean solo académicamente interesantes, sino que puedan también guiar los esfuerzos reales con los que lidiar con los desafíos planteados por lo que resulta ser inteligencia artificial y *poder* artificial: la IA es tanto tecnológica como *política*.

Por otra parte, junto con este uso práctico, este ejercicio de *filosofía política aplicada* de la IA también tiene importancia filosófica que va más allá de su aplicación: resultó que la conceptualización de la política de la IA y de la robótica no es una cuestión de aplicar simplemente las nociones existentes de la filosofía política y de la teoría social, sino que *nos invita a cuestionar los conceptos y valores mismos* (libertad, igualdad, justicia, democracia, poder, política antropocéntrica) y a revisar cuestiones interesantes sobre la naturaleza y el futuro de la política. Por ejemplo, ¿cuál es y cuál debería ser el papel del conocimiento experto, la racionalidad y las emociones en la política? Y ¿qué significa una política posantropocéntrica, una vez que nos cuestionamos la posición central y hegemónica de los humanos? El libro muestra que las discusiones sobre la IA nos invitan, en algún sentido nos «fuerzan», a revisar conceptos y discusiones filosófico-políticos, y que, al final, también nos desafían a cuestionar al ser humano y el humanismo; o, al menos, algunas versiones problemáticas de estos conceptos.

A la luz de la experiencia de que reflexionar sobre la tecnología nos inquieta y a veces desestabiliza nuestro pensamiento sobre la política, sugiero que la filosofía política del siglo XXI ya no puede hacerse, ni se debe seguir haciendo, sin responder a la cuestión relacionada con la tecnología. Tenemos que *zusammendenken* («reflexionar de manera conjunta sobre») la política y la tecnología: reflexionar sobre una de las partes no es posible sin reflexionar sobre la otra. Se necesita urgentemente más diálogo entre estos

campos de reflexión y, quizás, en última instancia, debería haber una fusión de los dos.

Y ya es hora de que lo hagamos. Como toda filosofía, de acuerdo con Hegel, quien en 1820 afirmó en sus *Principios de la filosofía del derecho* que la filosofía es «su propio tiempo aprehendido en pensamientos» (Hegel, 2008, 15), la filosofía política debería reaccionar y reflejar su tiempo; y quizás no pueda hacer nada más que eso: no puede realmente trascender su tiempo. Parafraseando la máxima latina citada por Hegel, aquí es donde tenemos que saltar: es el momento para un *salto mortale*. Y nuestro tiempo no es solo un tiempo de malestar y transformación sociales, medioambientales y existenciales, es también un tiempo en el que nuevas *tecnologías*, como la IA, están estrechamente enmarañadas con estos cambios y desarrollos. Es el tiempo de la IA. Reflexionar sobre el futuro de la política necesita, entonces, estar conectado fundamentalmente con la reflexión sobre las tecnologías y su relación con la política. En este caso: la IA está aquí, la IA es (de) nuestro tiempo y, por tanto, la *IA es allí donde tenemos que saltar y reflexionar*. Basándonos en la filosofía política y en la teorización relevante relacionada con ella (por ejemplo, teoría social sobre el poder, teoría poshumanista), este libro ha proporcionado alguna guía para este salto y reflexión.

Sin embargo, esto es simplemente el comienzo, un primer paso o, si se quiere, un prolegómeno: una introducción crítica a un proyecto más amplio que concierne a la política y a la IA y, de forma más general, a la política y a la tecnología. Terminaré este libro ofreciendo una idea del trabajo que queda por delante.

QUÉ SE NECESITA HACER A CONTINUACIÓN: LA CUESTIÓN RESPECTO DE LAS TECNOLOGÍAS POLÍTICAS

Como introducción y como libro de filosofía, el foco de este volumen se ha puesto en hacer preguntas más bien que en proporcionar respuestas. Ha ofrecido sugerencias sobre cómo puede ayudar a esta empresa la filosofía política, sugerencias que han crecido hasta llegar a ser una caja de herramientas y un marco, una estructura que ayude a la hora de discutir la

política de la IA. Sin embargo, se necesita más trabajo. Más precisamente, hay al menos dos tipos de pasos adicionales necesarios.

En primer lugar, se necesita más investigación y reflexión para continuar con el desarrollo del marco conceptual. Es como un andamio: sirve de apoyo, pero es temporal, y lo que se necesita ahora es continuar la construcción. Como los escritos sobre la política de la IA están creciendo rápidamente, no tengo duda de que habrá más elaboraciones sobre asuntos tales como el sesgo en la IA, el poder de los gigantes tecnológicos, la IA y la democracia, etc. Pero, a la luz del proyecto de este libro, que consiste en ayudar a dar a luz a una *filosofía política* de la IA, espero especialmente que (1) más *filósofos* escriban sobre la política de la IA (actualmente, a menudo lo hacen personas de otras disciplinas y existen muchos escritos no académicos que hacen poco más que arañar la superficie) y que (2) se haga más trabajo usando la *filosofía política* y la teoría social, cuyos recursos están actualmente infrautilizados en la reflexión normativa sobre la IA y la filosofía de la tecnología, quizás porque son menos familiares o populares que la ética. Como Langdon Winner, un importante filósofo de la tecnología, ya había defendido en la década de los 80 del siglo pasado, la tecnología es política. Winner advirtió de que, en lugar de lograr más democratización e igualdad social, las nuevas tecnologías bien podían conducir a más poder para aquellos que ya tenían bastante (Winner, 1986, 107). Como muestra mi libro, usando los recursos de la filosofía política, podemos continuar desarrollando la idea de que la tecnología es política y discutir críticamente el impacto de tecnologías como la IA.

En segundo lugar, si la política se concibe como algo que es, por definición, de interés público y en la que todos deberíamos participar, reflexionar sobre la política de la IA debería también tener lugar fuera de la academia y debería hacerse por todo tipo de partes interesadas en todo tipo de contextos. La política de la IA no es algo sobre lo que solo deberíamos pensar y escribir en libros: es algo que deberíamos también *hacer*. Si rechazamos la idea platónica de que solo los filósofos y los expertos deberían gobernar, entonces descubrir qué es una buena sociedad a la luz de la IA es algo que deberíamos descubrir juntos: la política de la IA debería discutirse públicamente y de forma participativa, y debería hacerse de una

manera inclusiva. Pero esto no excluye el papel de la filosofía y de los filósofos: los conceptos y teorías de la filosofía política, como los ofrecidos en este libro, pueden ayudar a elevar la calidad de tales discusiones públicas. Por ejemplo, actualmente se dice a menudo que la IA amenaza la democracia, pero no queda claro por qué y qué se quiere decir con «democracia». Como he mostrado en el capítulo 4, la filosofía política, con ayuda de la filosofía de la tecnología y de los medios de comunicación, puede ayudar a aclarar este tema. Además, dados algunos de los peligros indicados en este libro (por ejemplo, los sesgos y las formas de discriminación en la sociedad, las cámaras de eco y los filtros burbuja a través de las redes sociales, el peligro del uso totalitario de la IA), nos enfrentamos al reto de *cómo* abordar estos desafíos y de cómo mejorar la discusión. ¿Qué procedimientos, infraestructuras y formas de conocimiento necesitamos para una discusión democrática e inclusiva sobre la IA y otras tecnologías? Y, ciertamente, ¿qué *tecnologías* (no) necesitamos y cómo (no) es mejor usarlas? ¿Qué tipo de redes sociales necesitamos y cuál debería ser el papel y lugar de la IA, si es que tiene alguno? Reflexionar sobre cómo hacer la política de la IA de una forma democrática e inclusiva nos lleva de nuevo a una de las preguntas básicas que debemos hacernos sobre la democracia y la política misma: *cómo* hacerla. Y, si la pregunta relacionada con la política y la pregunta relacionada con la tecnología se encuentran realmente conectadas, entonces esa pregunta puede también formularse como: ¿qué *tecnologías políticas* necesitamos y queremos?

Finalmente, en la medida en que este libro ha usado, y respondido, a la filosofía política anglosajona estándar, el marco presentado ha tomado prestado también parcialmente sus sesgos, orientaciones político-culturales y limitaciones. Por ejemplo, muchas discusiones que he tenido durante mi viaje a lo largo del panorama filosófico-político de habla inglesa simplemente dan por hecho el contexto y cultura políticos de los Estados Unidos, ignorando, por lo tanto, otros enfoques y contextos en otras partes del mundo y (seguramente peor) ignorando *cómo sus opiniones filosóficas, argumentos y suposiciones están conformados por su propio contexto político-cultural*. Además, la mayor parte de la filosofía política moderna se centra en el contexto del Estado nación y no logra abordar los desafíos que

surgen en un contexto global. Cuando se trata de continuar el desarrollo de la reflexión sobre la IA y la política, tanto a nivel académico como no académico, será vital abordar estas cuestiones y recorrer los desafíos planteados por la política de la IA de formas que tomen en consideración el contexto *global* en el que surgen y crear la sensibilidad suficiente a las *diferencias culturales* con respecto a la forma en la que personas y culturas diferentes reflexionan sobre la tecnología, sobre la política y, ciertamente, sobre los humanos. Primero, dado que la IA no entiende de fronteras y tiene un impacto que va más allá de las naciones Estado, y dado que existen en el mundo muchos actores diferentes en el ámbito de la IA (no solo los Estados Unidos, sino también Europa, China, etc.), es importante reflexionar sobre la política de la IA en un contexto global y, quizás, desarrollar una política global de la IA. Tal proyecto plantea sus propios desafíos. Por ejemplo, varias organizaciones intergubernamentales y no gubernamentales ya están desarrollando políticas para lidiar con la IA. Pero ¿es este trabajo *internacional* (entre naciones) suficiente o necesitamos (más) formas *supranacionales* de gobernanza? ¿Necesitamos nuevas instituciones políticas, nuevas tecnologías políticas para gobernar la IA a nivel global? ¿*Qué tecnologías políticas globales necesitamos?* Segundo, debe tenerse en cuenta que los textos de los que se hace uso en este libro reflejan contextos políticos específicos. Por ejemplo, cuando Eubanks (2018) critica una forma moralista específica de lidiar con la pobreza, esta crítica está situada dentro de la cultura política estadounidense, que no es necesariamente compartida por otros países y que tiene sus propios desafíos relacionados con *esa* cultura política. Reflexionar sobre la política de la IA necesitará ser algo más sensible de cara a la dimensión de la diferencia cultural, especialmente si busca convertirse en más relacional, más imaginativa, más responsable y más relevante en la práctica, tanto en el contexto local como en el global.

Para concluir, este libro incorpora no solo una recomendación de usar la filosofía política para reflexionar sobre la IA. De forma más general y más ambiciosa, es también una invitación a correr el riesgo de este *salto mortal* de pensar juntos sobre la política y la tecnología y hacerlo de una forma que responda a lo que está ocurriendo en nuestras sociedades y en el mundo. Es

el momento de hacerlo, y es muy necesario. Si no tomamos este camino, no lograremos suficiente distancia crítica y reflexiva frente a lo que tecnologías como la IA ya nos están haciendo a nosotros y a la política y seremos víctimas indefensas de la inteligencia y del poder artificiales. Esto es, seremos las víctimas indefensas de nosotros mismos y de nuestra sociedad: de las propias tecnologías humanas (demasiado humanas), las metáforas, los dualismos y las estructuras de poder que nosotros permitimos que nos gobiernen. Esto evoca las narrativas distópicas y, desafortunadamente, los ejemplos del mundo real que aparecen una y otra vez a lo largo de este libro: historias y casos que muestran cómo los principios y valores políticos clave se encuentran bajo amenaza. Podemos, y debemos, hacerlo mejor. Reflexionar sobre las tecnologías políticas, conectar la reflexión sobre la tecnología con el cuestionamiento de los principios y estructuras fundamentales de nuestras sociedades y del orden político global nos puede ayudar a crear y a vivir historias mejores y más positivas: no sobre un futuro distante, sino en el aquí y el ahora. Historias sobre la IA, sobre nosotros y sobre otros seres y cosas que importan.

Bibliografía⁴

- AAVITSLAND, V. L. (2019), «The Failure of Judgment: Disgust in Arendt's Theory of Political Judgment», *Journal of Speculative Philosophy* 33 (3), págs. 537-550.
- ADORNO, T. (1983), *Prisms*, trad. por S. Weber y S. Weber, Cambridge, MA, MIT Press. [Disponible en castellano como *Prismas* (1962), Barcelona, Ariel.]
- AGAMBEN, G. (1998), *Homo Sacer: Sovereign Power and Bare Life*, trad. por D. Heller-Roazen, Stanford, Stanford University Press. [Disponible en castellano como *Homo sacer: El poder soberano y la nuda vida* (2010), Valencia, Pre-textos.]
- AI INSTITUTE (2019), «AI and Climate Change: How They're Connected, and What We Can Do about It», *Medium*, 17 de octubre. [Disponible en: <<https://medium.com/@AINowInstitute/ai-and-climate-change-how-theyre-connected-and-what-we-can-do-about-it-6aa8d0f5b32c>>.]
- ALAIMO, S. (2016), *Exposed: Environmental Politics and Pleasures in Posthuman Times*, Mineápolis, University of Minnesota Press.
- ALBRECHTSLUND, A. (2008), «Online Social Networking as Participatory Surveillance», *First Monday* 13(3). [Disponible en: <<https://doi.org/10.5210/fm.v13i3.2142>>.]
- ANDREJEVIC, M. (2020), *Automated Media*, Nueva York, Routledge, Arendt, H. (1943), «We Refugees», *Menorah Journal* 31(1), págs. 69-77.
- ARENDT, H. (1958), *The Human Condition*, Chicago, University of Chicago Press. [Disponible en castellano como *La condición humana* (2005), Barcelona, Paidós Ibérica.]
- (1968), *Between Past and Future*, Nueva York, Viking Press. [Disponible en castellano como *Entre el pasado y el futuro* (2016), Barcelona, Ariel.]
- (2006), *Eichmann in Jerusalem: A Report on the Banality of Evil*, Nueva York, Penguin. [Disponible en castellano como *Eichmann en Jerusalén*

- (2021), Barcelona, Debolsillo.]
- (2017), *The Origins of Totalitarianism*, Londres, Penguin. [Disponible en castellano como *Los orígenes del totalitarismo* (2006), Madrid, Alianza.]
- ASDAL, K.; DRUGLITRØ, T. y HINCHLIFFE, S. (2017), «Introduction: The “More-Than-Human” Condition», en K. Asdal, T. Druglitrø y S. Hinchliffe (eds.), *Humans, Animals, and Biopolitics*, Abingdon, Routledge, págs. 1-29.
- ATANASOSKI, N. y VORA, K. (2019), *Surrogate Humanity: Race, Robots, and the Politics of Technological Futures*, Durham, NC, Duke University Press.
- AUSTIN, J. L. (1962), *How to Do Things with Words*, Cambridge, MA, Harvard University Press.
- AZMANOVA, A. (2020), *Capitalism on Edge: How Fighting Precarity Can Achieve Radical Change without Crisis or Utopia*, Nueva York, Columbia University Press.
- BAKARDJIEVA, M. y GADEN, G. (2011), «Web 2.0 Technologies of the Self», *Philosophy & Technology* 25, págs. 399-413.
- (2003), «Posthumanist Performativity: Towards an Understanding of How Matter Comes to Matter», *Signs: Journal of Women in Culture and Society* 28(3), págs. 801-831.
- (2007), *Meeting the Universe Halfway: Quantum Physics and the Entanglement of Matter and Meaning*, Durham, NC, Duke University Press.
- (2015), «Transmaterialities: Trans*Matter/Realities and Queer Political Imaginings», *GLQ: A Journal of Lesbian and Gay Studies* 21(2-3), págs. 387-422.
- BARTNECK, C.; LÜTGE, C.; WAGNER, A. y WELSH, S. (2021), *An Introduction to Ethics in Robotics and AI*, Cham, Springer.
- BARTOLETTI, I. (2020), *An Artificial Revolution: On Power, Politics and AI*, Londres, The Indigo Press. BBC (2018), «Fitbit Data Used to Charge US Man with Murder», *BBC News*, 4 de octubre. [Disponible en: <<https://www.bbc.com/news/technology-45745366>>.]
- BELL, D. A. (2016), *The China Model: Political Meritocracy and the Limits of Democracy*, Princeton, Princeton University Press.
- BENJAMIN, R. (2019a), *Race After Technology*, Cambridge, Polity.

- (2019b), *Captivating Technology: Race, Carceral Technoscience, and Liberatory Imagination in Everyday Life*, Durham, NC, Duke University Press.
- BERARDI, F. (2017), *Futurability: The Age of Impotence and the Horizon of Possibility*, Londres, Verso. [Disponible en castellano como *Futurabilidad: la era de la impotencia y el horizonte de la posibilidad* (2019), Buenos Aires, Caja Negra.]
- BERLIN, I. (1997), «Two Concepts of Liberty», en I. Berlin, *The Proper Study of Mankind*, Londres, Chatto & Windus, págs. 191-242. [Disponible en castellano como *Dos conceptos de libertad. El fin justifica los medios. Mi trayectoria intelectual* (2014), Madrid, Alianza.]
- BERMAN, J. (2011), «Futurist Ray Kurzweil Says He Can Bring His Dead Father Back to Life Through a Computer Avatar», *ABC News*, 10 de agosto. [Disponible en: <<https://abcnews.go.com/Technology/futurist-ray-kurzweil-bring-dead-father-back-life/story?id=14267712>>.]
- BERNAL, N. (2020), «They Claim Uber's Algorithm Fired Them. Now They're Taking It to Court», *Wired*, 2 de noviembre. [Disponible en: <<https://www.wired.co.uk/article/uber-fired-algorithm>>.]
- BIETTI, E. (2020), «Consent as a Free Pass: Platform Power and the Limits of Information Turn», *Pace Law Review* 40(1), págs. 310-398.
- BINNS, R. (2018), «Fairness in Machine Learning: Lessons from Political Philosophy», Actas de la primera conferencia sobre justicia, responsabilidad y transparencia. *Proceedings of Machine Learning Research* 81, págs. 149-159. [Disponible en: <<http://proceedings.mlr.press/v81/binns18a.html>>.]
- BIRHANE, A. (2020), «Algorithmic Colonization of Africa», *SCRIPTed: A Journal of Law, Technology, & Society* 17(2). [Disponible en: <<https://script-ed.org/article/algorithmic-colonization-of-africa/>>.]
- BLOOM, P. (2019), *Monitored: Business and Surveillance in a Time of Big Data*, Londres, Pluto Press.
- BODDINGTON, P. (2017), *Towards a Code of Ethics of Artificial Intelligence*, Cham, Springer.
- BOSTROM, N. (2014), *Superintelligence: Paths, Dangers, Strategies*, Oxford, Oxford University Press. [Disponible en castellano como *Superinteligencia. Caminos, peligros, estrategias* (2016), Zaragoza, Teell.]

- BOSTROM, N. y YUDKOWSKY, E. (2014), «The Ethics of Artificial Intelligence», en K. Frankish y W. Ramsey (eds.), *Cambridge Handbook of Artificial Intelligence*, Nueva York, Cambridge University Press, págs. 316-334.
- BOURDIEU, P. (1990), *The Logic of Practice*, trad. por R. Nice, Stanford, Stanford University Press. [Disponible en castellano como *El sentido práctico* (1991), Barcelona, Taurus.]
- BOZDAG, E. (2013), «Bias in Algorithmic Filtering and Personalization», *Ethics and Information Technology* 15(3), págs. 209-227.
- BRADLEY, A. (2011), *Originary Technicity: The Theory of Technology from Marx to Derrida*, Basingstoke, Palgrave Macmillan.
- BRAIDOTTI, R. (2016), «Posthuman Critical Theory», en D. Banerji y M. Paranjape (eds.), *Critical Posthumanism and Planetary Futures*, Nueva Delhi, Springer, págs. 13-32.
- (2017), «Posthuman Critical Theory», *Journal of Posthuman Studies* 1(1), págs. 9-25.
- (2020), «“We” Are in This Together, but We Are Not One and the Same», *Journal of Bioethical Inquiry* 17(4), págs. 465-469.
- BROUSSARD, M. (2019), *Artificial Unintelligence: How Computers Misunderstand the World*, Cambridge, MA, MIT Press.
- BRYSON, J. J. (2010), «Robots Should Be Slaves», en Y. Wilks (ed.), *Close Engagements with Artificial Companions*, Ámsterdam, John Benjamins Publishing, págs. 63-74.
- BUTLER, J. (1988), «Performative Acts and Gender Constitution: An Essay in Phenomenology and Feminist Theory», *Theatre Journal* 40(4), págs. 519-531.
- (1989), «Foucault and the Paradox of Bodily Inscriptions», *Journal of Philosophy* 86(11), págs. 601-607.
- (1993), *Bodies That Matter: On the Discursive Limits of «Sex»*, Londres, Routledge. [Disponible en castellano como *Cuerpos que importan: sobre los límites materiales y discursivos del sexo* (2003), Buenos Aires, Paidós Ibérica.]
- (1997), *Excitable Speech: A Politics of the Performative*, Nueva York, Routledge.
- (1999), *Gender Trouble: Feminism and the Subversion of Identity*, Nueva York, Routledge. [Disponible en castellano como *El género en*

- disputa: el feminismo y la subversión de la identidad* (2007), Buenos Aires, Paidós Ibérica.]
- (2004), *Precarious life: The Powers of Mourning and Violence*, Londres, Verso. [Disponible en castellano como *Vida precaria: el poder del duelo y la violencia* (2007), Buenos Aires, Paidós Ibérica.]
- CALISKAN, A.; BRYSON, J. J. y NARAYANAN, A. (2017), «Semantics Derived Automatically from Language Corpora Contain Human-Like Biases», *Science* 356(6334), págs. 183-186.
- CALLICOTT, J. B. (1989), *In Defense of the Land Ethic: Essays in Environmental Philosophy*, Albany, State University of New York Press.
- CANAVAN, G. (2015), «Capital as Artificial Intelligence», *Journal of American Studies* 49(4), págs. 685-709.
- CASTELLS, M. (2001), *The Internet Galaxy: Reflections on the Internet, Business, and Society*, Oxford, Oxford University Press. [Disponible en castellano como *La galaxia internet* (2007), Barcelona, Plaza & Janés.]
- CELERMAJER, D.; SCHLOSBERG, D.; RICKARDS, L.; STEWART-HARAWIRA, M.; THALER, M.; TSCHAKERT, P.; VERLIE, B. y WINTER, C. (2021), «Multispecies Justice, Theories, Challenges, and a Research Agenda for Environmental Politics», *Environmental Politics* 30(1-2), págs. 119-140.
- CHENEY-LIPPOLD, J. (2017), *We Are Data: Algorithms and the Making of Our Digital Selves*, Nueva York, New York University Press.
- CHOU, M.; MOFFITT, B. y BRYANT, O. (2020), *Political Meritocracy and Populism: Cure or Curse?*, Nueva York, Routledge.
- CHRISTIANO, T. (ed.) (2003), *Philosophy and Democracy: An Anthology*, Oxford, Oxford University Press.
- CHRISTIANO, T. y BAJAJ, S. (2021), «Democracy», *Stanford Encyclopedia of Philosophy*. [Disponible en: <https://plato.stanford.edu/entries/democracy/>.]
- CHRISTMAN, J. (2004), «Relational Autonomy, Liberal Individualism, and the Social Constitution of Selves», *Philosophical Studies* 117(1-2), págs. 143-164.
- COECKELBERGH, M. (2009a), «The Public Thing: On the Idea of a Politics of Artefacts», *Techné* 13(3), págs. 175-181.
- (2009b), «Distributive Justice and Cooperation in a World of Humans and Non-Humans: A Contractarian Argument for Drawing Non-Humans into the Sphere of Justice», *Res Publica* 15(1), págs. 67-84.

- (2012), *Growing Moral Relations: Critique of Moral Status Ascription*, Basingstoke y Nueva York, Palgrave Macmillan.
 - (2013), *Human Being @ Risk*, Dordrecht, Springer.
 - (2014), «The Moral Standing of Machines: Towards a Relational and Non-Cartesian Moral Hermeneutics», *Philosophy & Technology* 27(1), págs. 61-77.
 - (2015a), «The Tragedy of the Master: Automation, Vulnerability, and Distance», *Ethics and Information Technology* 17(3), págs. 219-229.
 - (2015b), *Environmental Skill*, Abingdon Routledge.
 - (2017), «Beyond “Nature”. Towards More Engaged and Care-Full Ways of Relating to the Environment», en H. Kopnina y E. Shoreman-Ouimet (eds.), *Routledge Handbook of Environmental Anthropology*, Abingdon, Routledge, págs. 105-116.
 - (2019a), *Introduction to Philosophy of Technology*, Nueva York, Oxford University Press.
 - (2019b), *Moved by Machines: Performance Metaphors and Philosophy of Technology*, Nueva York, Routledge.
 - (2019c), «Technoperformances: Using Metaphors from the Performance Arts for a Postphenomenology and Posthermeneutics of Technology Use», *AI & Society* 35(3), págs. 557-568.
 - (2020), *AI Ethics*, Cambridge, MA, MIT Press. [Disponible en castellano como *Ética de la inteligencia artificial* (2021), Madrid, Cátedra.]
 - (2021), «How to Use Virtue Ethics for Thinking about the Moral Standing of Social Robots: A Relational Interpretation in Terms of Practices, Habits, and Performance», *International Journal of Social Robotics* 13(1), págs. 31-40.
- CONFAVREUX, J. y RANCIÈRE, J. (2020), «The Crisis of Democracy», Verso, 24 de febrero. [Disponible en: <<https://www.versobooks.com/blogs/4576-jacques-ranciere-the-crisis-of-democracy>>.]
- COOK, G.; LEE, J.; TSAI, T.; KONG, A.; DEANS, J.; JOHNSON, B. y JARDIN, E. (2017), *Clicking Clean: Who Is Winning the Race to Build a Green Internet?*, Washington, Greenpeace.
- COTTER, K. y REISDORF, B. C. (2020), «Algorithmic Knowledge Gaps: A New Dimension of (Digital) Inequality», *International Journal of Communication* 14, págs. 745-765.

- COULDRY, N.; LIVINGSTONE, S. y MARKHAM, T. (2007), *Media Consumption and Public Engagement: Beyond the Presumption of Attention*, Nueva York, Palgrave Macmillan.
- COULDRY, N. y MEJIAS, U. A. (2019), *The Costs of Connection: How Data Is Colonizing Human Life and Appropriating It for Capitalism*, Stanford, Stanford University Press.
- CRARY, J. (2014), *24/7, Late Capitalism and the Ends of Sleep*, Londres, Verso. [Disponible en castellano como *24/7: capitalismo tardío y el fin del sueño* (2015), Barcelona, Ariel.]
- CRAWFORD, K. (2021), *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*, New Haven, Yale University Press.
- CRAWFORD, K. y CALO, R. (2016), «There Is a Blind Spot in AI Research», *Nature* 538, págs. 311-313.
- CRIADO PEREZ, C. (2019), *Invisible Women: Data Bias in a World Designed for Men*, Nueva York, Abrams Press. [Disponible en castellano como *La mujer invisible* (2020), Barcelona, Seix Barral.]
- CRUTZEN, P. (2006), «The “Anthropocene”», en E. Ehlers y T. Krafft (eds.), *Earth System Science in the Anthropocene*, Berlín, Springer, págs. 13-18.
- CUDWORTH, E. y HOBDEN, S. (2018), *The Emancipatory Project of Posthumanism*, Londres, Routledge.
- CURRY, P. (2011), *Ecological Ethics. An Introduction*, 2.^a ed., Cambridge, Polity.
- DAHL, R. A. (2006), *A Preface to Democratic Theory*, Chicago, University of Chicago Press.
- DAMNJANOVIĆ, I. (2015), «Polity without Politics? Artificial Intelligence versus Democracy, Lessons from Neal Asher’s Polity Universe», *Bulletin of Science, Technology & Society* 35(3-4), págs. 76-83.
- DANAHER, J. (2020), «Welcoming Robots into the Moral Circle: A Defence of Ethical Behaviorism», *Science and Engineering Ethics* 26(4), págs. 2023-2049.
- DARLING, K. (2016), «Extending Legal Protection to Social Robots: The Effects of Anthropomorphism, Empathy, and Violent Behavior towards Robotic Objects», en R. Calo, A. M. Froomkin e I. Kerr (eds.), *Robot Law*, Cheltenham, Edward Elgar Publishing, págs. 213-232.
- DAUVERGNE, P. (2020), «The Globalization of Artificial Intelligence: Consequences for the Politics of Environmentalism», *Globalizations*

18(2), págs. 285-299.

DEAN, J. (2009), *Democracy and Other Neoliberal Fantasies: Communicative Capitalism and Left Politics*, Durham, NC, Duke University Press.

DELEUZE, G. y GUATTARI, F. (1987), *A Thousand Plateaus: Capitalism and Schizophrenia*, trad. por B. Massumi, Mineápolis, University of Minnesota Press. [Disponible en castellano como *Mil mesetas* (1994), Valencia, Pre-textos.]

DENT, N. (2005), *Rousseau*, Londres, Routledge.

DEPLAZES, A. y HUPPENBAUER, M. (2009), «Synthetic Organisms and Living Machines», *Systems and Synthetic Biology* 3(55). [Disponible en: <<https://doi.org/10.1007/s11693-009-9029-4>>.]

DERRIDA, J. (1976), *Of Grammatology*, trad. por G. C. Spivak. Baltimore, MD, Johns Hopkins University Press. [Disponible en castellano como *De la gramatología* (1986), México, Siglo XXI.]

— (1981), «Plato's Pharmacy», en J. Derrida, *Dissemination*, trad. por B. Johnson, Chicago, University of Chicago Press, págs. 63-171. [Disponible en castellano como *La diseminación* (2015), Madrid, Fundamentos.]

DETROW, S. (2018), «What Did Cambridge Analytica Do during the 2016 Election?» *NPR*, 21 de marzo. [Disponible en: <<https://text.npr.org/595338116>>.]

DEWEY, J. (2001), *Democracy and Education*, Hazleton, PA, Penn State Electronic Classics Series. [Disponible en castellano como *Democracia y educación: una introducción a la filosofía de la educación* (1995), Madrid, Morata.]

DIAMOND, L. (2019), «The Threat of Postmodern Totalitarianism», *Journal of Democracy* 30(1), págs. 20-24.

DIGNUM, V. (2019), *Responsible Artificial Intelligence*, Cham, Springer.

DIXON, S. (2007), *Digital Performance: A History of New Media in Theater, Dance, Performance Art, and Installation*, Cambridge, MA, MIT Press.

DJEFFAL, C. (2019), «AI, Democracy and the Law», en A. Sudmann (ed.), *The Democratization of Artificial Intelligence: Net Politics in the Era of Learning Algorithms*, Bielefeld, Transcripción, págs. 255-283.

DONALDSON, S. y KYMLICKA, W. (2011), *Zoopolis: A Political Theory of Animal Rights*, Nueva York, Oxford University Press. [Disponible en

- castellano como *Zoópolis: una revolución animalista* (2018), Madrid, Errata.]
- DOWNING, L. (2008), *The Cambridge Introduction to Michel Foucault*, Nueva York, Cambridge University Press.
- DUBBER, M.; PASQUALE, F. y DAS, S. (2020), *The Oxford Handbook of Ethics of AI*, Oxford, Oxford University Press.
- DWORKIN, R. (2011), *Justice for Hedgehogs*, Cambridge, MA, Belknap Press. [Disponible en castellano como *Justicia para erizos* (2016), Madrid, Fondo de Cultura Económica.]
- (2020), «Paternalism», *Stanford Encyclopedia of Philosophy*. [Disponible en: <<https://plato.stanford.edu/entries/paternalism/>>.]
- DYER-WITHEFORD, N. (1999), *Cyber-Marx: Cycles and Circuits of Struggle in High-Technology Capitalism*, Urbana, University of Illinois Press.
- (2015), *Cyber-Proletariat Global Labour in the Digital Vortex*, Londres, Pluto Press.
- DYER-WITHEFORD, N., KJØSEN, A. M. y STEINHOFF, J. (2019), *Inhuman Power: Artificial Intelligence and the Future of Capitalism*, Londres, Pluto Press.
- EL-BERMAWY, M. M. (2016), «Your Filter Bubble Is Destroying Democracy», *Wired*, 18 de noviembre. [Disponible en: <<https://www.wired.com/2016/11/filter-bubble-destroying-democracy/>>.]
- ELKIN-KOREN, N. (2020), «Contesting Algorithms: Restoring the Public Interest in Content Filtering by Artificial Intelligence», *Big Data & Society* 7(2). [Disponible en: <<https://doi.org/10.1177/2053951720932296>>.]
- ERIKSSON, K. (2012), «Self-Service Society: Participative Politics and New Forms of Governance», *Public Administration* 90(3), págs. 685-698.
- ESHUN, K. (2003), «Further Considerations of Afrofuturism», *CR: The New Centennial Review* 3(2), págs. 287-302.
- ESTLUND, D. (2008), *Democratic Authority: A Philosophical Framework*, Princeton, Princeton University Press. [Disponible en castellano como *La autoridad democrática* (2013), Madrid, Siglo XXI.]
- EUBANKS, V. (2018), *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*, Nueva York, St. Martin's Press. [Disponible en castellano como *La automatización de la desigualdad*:

- herramientas de tecnología avanzada para supervisar y castigar a los pobres* (2021), Madrid, Capitán Swing.]
- FARKAS, J. (2020), «A Case against the Post-Truth Era: Revisiting Mouffe's Critique of Consensus-Based Democracy», en M. Zimdars y K. McLeod (eds.), *Fake News: Understanding Media and Misinformation in the Digital Age*, Cambridge, MA, MIT Press, págs. 45-54.
- FARKAS, J. y SCHOU, J. (2018), «Fake News as a Floating Signifier: Hegemony, Antagonism and the Politics of Falsehood», *Javnost-The Public* 25(3), págs. 298-314.
- (2020), *Post-Truth, Fake News and Democracy: Mapping the Politics of Falsehood*, Nueva York, Routledge.
- FEENBERG, A. (1991), *Critical Theory of Technology*, Oxford, Oxford University Press.
- (1999), *Questioning Technology*, Londres, Routledge.
- FERRANDO, F. (2019), *Philosophical Posthumanism*, Londres, Bloomsbury Academic.
- FLORIDI, L. (2013), *The Ethics of Information*, Oxford, Oxford University Press.
- (2014), *The Fourth Revolution*, Oxford, Oxford University Press.
- (2017), «Roman Law Offers a Better Guide to Robot Rights Than Sci-Fi», *Financial Times*, 22 de febrero. [Disponible en: <https://www.academia.edu/31710098/Roman_law_offers_a_better_guide_to_robot_rights_than_sci-fi>.]
- FLORIDI, L. y SANDERS, J. W. (2004), «On the Morality of Artificial Agents», *Minds & Machines* 14(3), págs. 349-379.
- FOGG, B. (2003), *Persuasive Technology: Using Computers to Change What We Think and Do*. San Francisco, Morgan Kaufmann.
- FORD, M. (2015), *The Rise of the Robots: Technology and the Threat of a Jobless Future*, Nueva York, Basic Books. [Disponible en castellano como *El auge de los robots: la tecnología y la amenaza de un futuro sin empleo* (2016), Barcelona, Paidós Ibérica.]
- FOUCAULT, M. (1977), *Discipline and Punish: The Birth of the Prison*, trad. por A. Sheridan, Nueva York, Vintage Books. [Disponible en castellano como *Vigilar y castigar* (2012), Capital Federal, Siglo XXI.]
- (1980), *Power/Knowledge: Selected Interviews and Other Writings 1972-1977*. Editado por C. Gordon, trad. por C. Gordon, L. Marshall, J.

- Mepham y K. Soper, Nueva York, Pantheon Books.
- (1981), *History of Sexuality: Volume 1: An Introduction*, trad. por R. Hurley, Londres, Penguin. [Disponible en castellano como *Historia de la sexualidad 1: la voluntad de saber* (2006), Madrid, Siglo XXI.]
- (1988), «Technologies of the Self», en L. H. Martin, H. Gutman y P. H. Hutton (eds.), *Technologies of the Self: A Seminar with Michel Foucault*, Amherst, University of Massachusetts Press, págs. 16-49.
- FRANKFURT, H. (2000), «Distinguished Lecture in Public Affairs: The Moral Irrelevance of Equality», *Public Affairs Quarterly* 14(2), págs. 87-103.
- FRANKFURT, H. (2015), *On Inequality*, Princeton, Princeton University Press. [Disponible en castellano como *Sobre la desigualdad* (2016), Barcelona, Paidós Ibérica.]
- FUCHS, C. (2014), *Social Media: A Critical Introduction*, Londres, Sage Publications.
- (2020), *Communication and Capitalism: A Critical Theory*, Londres, University of Westminster Press.
- FUCHS, C.; BOERSMA, K.; ALBRECHTSLUND, A. y SANDOVAL, M. (eds.) (2012), *Internet and Surveillance: The Challenges of Web 2.0 and Social Media*, Londres, Routledge.
- FUKUYAMA, F. (2006), «Identity, Immigration, and Liberal Democracy», *Journal of Democracy* 17(2), págs. 5-20.
- (2018a), «Against Identity Politics: The New Tribalism and the Crisis of Democracy», *Foreign Affairs* 97(5), págs. 90-115.
- FUKUYAMA, F. (2018b), *Identity: The Demand for Dignity and the Politics of Resentment*, Nueva York, Farrar, Straus y Giroux. [Disponible en castellano como *Identidad: la demanda de dignidad y las políticas de resentimiento* (2019), Barcelona, Deusto ediciones.]
- GABRIELS, K. y COECKELBERGH, M. (2019), «Technologies of the Self and the Other, How Self-Tracking Technologies Also Shape the Other», *Journal of Information, Communication and Ethics in Society* 17(2). [Disponible en: <<https://doi.org/10.1108/JICES-12-2018-0094>>.]
- GARNER, R. (2003), «Animals, Politics, and Justice: Rawlsian Liberalism and the Plight of Non-Humans», *Environmental Politics* 12(2), págs. 3-22.
- (2012), «Rawls, Animals and Justice: New Literature, Same Response», *Res Publica* 18(2), págs. 159-172.

- GELLERS, J. C. (2020), «Earth System Governance Law and the Legal Status of Non-Humans in the Anthropocene», *Earth System Governance* 7. [Disponible en: <<https://doi.org/10.1016/j.esg.2020.100083>>.]
- GIEBLER, H. y MERKEL, W. (2016), «Freedom and Equality in Democracies: Is There a Trade-Off?», *International Political Science Review* 37(5), págs. 594-605.
- GILLEY, B. (2016), «Technocracy and Democracy as Spheres of Justice in Public Policy», *Policy Sciences* 50(1), págs. 9-22.
- GITELMAN, L. y JACKSON, V. (2013), «Introduction», en L. Gitelman (ed.), «*Raw Data*» *Is an Oxymoron*, Cambridge, MA, MIT Press, págs. 1-14.
- Goodin, R. E. (2003), *Reflective Democracy*, Oxford, Oxford University Press.
- GORWA, R.; BINNS, R. y KATZENBACH, C. (2020), «Algorithmic Content Moderation: Technical and Political Challenges in the Automation of Platform Governance», *Big Data & Society* 7(1). [Disponible en: <<https://doi.org/10.1177/2053951719897945>>.]
- GRANKA, L. A. (2010), «The Politics of Search: A Decade Retrospective», *The Information Society Journal* 26(5), págs. 364-374.
- GRAY, C. H. (2000), *Cyborg Citizen: Politics in the Posthuman Age*, Londres, Routledge.
- GUNKEL, D. (2014), «A Vindication of the Rights of Machines», *Philosophy & Technology* 27(1), págs. 113-132.
- (2018), *Robot Rights*, Cambridge, MA, MIT Press.
- HABERMAS, J. (1990), *Moral Consciousness and Communicative Action*, trad. por C. Lenhart y S. W. Nicholson. Cambridge, MA, MIT Press. [Disponible en castellano como *Conciencia moral y acción comunicativa* (1996), Barcelona, Península.]
- HACKER, P. (2018), «Teaching Fairness to Artificial Intelligence: Existing and Novel Strategies against Algorithmic Discrimination under EU Law», *Common Market Law Review* 55(4), págs. 1143-1185.
- HAGGERTY, K. y ERICSON, R. (2000), «The Surveillant Assemblage», *British Journal of Sociology* 51(4), págs. 605-622.
- HAN, B. C. (2015), *The Burnout Society*, Stanford, Stanford University Press. [Disponible en castellano como *La sociedad del cansancio* (2012), Barcelona, Herder.]

- HARARI, Y. N. (2015), *Homo Deus: A Brief History of Tomorrow*, Londres, Harvill Secker. [Disponible en castellano como *Homo Deus: breve historia del mañana* (2016), Barcelona, Debate.]
- HARAWAY, D. (2000), «A Cyborg Manifesto», en D. Bell y B. M. Kennedy (eds.), *The Cybercultures Reader*, Londres, Routledge, págs. 291-324.
- (2003), *The Companion Species Manifesto: Dogs, People, and Significant Otherness*, Chicago, Prickly Paradigm Press. [Disponible en castellano como *Manifiesto de las especies de compañía* (2016), Álava, Sans soleil.]
- (2015), «Anthropocene, Capitalocene, Plantationocene, Chthulucene, Making Kin», *Environmental Humanities* 6, págs. 159-165.
- (2016), *Staying with the Trouble: Making Kin in the Chthulucene*, Durham, NC, Duke University Press. [Disponible en castellano como *Seguir con el problema: generar parentesco en el Chthuluceno* (2019), Bilbao, Consonni.]
- HARDT, M. (2015), «The Power to Be Affected», *International Journal of Politics, Culture, and Society* 28(3), págs. 215-222.
- HARDT, M. y NEGRI, A. (2000), *Empire*, Cambridge, MA, Harvard University Press. [Disponible en castellano como *Imperio* (2005), Barcelona, Paidós Ibérica.]
- HARVEY, D. (2019), *Marx, Capital and the Madness of Economic Reason*, Londres, Profile Books. [Disponible en castellano como *Marx, el capital y la locura de la razón económica* (2019), Madrid, Akal.]
- HEGEL, G. W. F. (1977), *Phenomenology of Spirit*, trad. por A. V. Miller. Oxford, Oxford University Press. [Disponible en castellano como *Fenomenología del espíritu* (2006), Valencia, Pre-textos.]
- (2008), *Outlines of the Philosophy of Right*, trad. por T. M. Knox. Oxford, Oxford University Press. [Disponible en castellano como *Principios de la filosofía del derecho* (1988), Barcelona, Edhasa.]
- (1977), *The Question Concerning Technology and Other Essays*, trad. por W. Lovitt, Nueva York, Garland Publishing. [Disponible en castellano como *La pregunta por la técnica* (2021), Barcelona, Herder.]
- HELBURG, N.; ESKENS, S., VAN DRUNEN, M.; BASTIAN, M. y MOELLER, J. (2019), «Implications of AI-Driven Tools in the Media for Freedom of Expression», Institute for Information Law (IViR). [Disponible en: <<https://rm.coe.int/coe-ai-report-final/168094ce8f>>.]

- HEYES, C. (2020), «Identity Politics», *Stanford Encyclopedia of Philosophy*.
[Disponible en: <<https://plato.stanford.edu/entries/identity-politics/>>.]
- HILDEBRANDT, M. (2015), *Smart Technologies and the End(s) of Law: Novel Entanglements of Law and Technology*, Cheltenham, Edward Elgar Publishing.
- HILDRETH, R.W. (2009), «Reconstructing Dewey on Power», *Political Theory* 37(6), págs. 780-807.
- HILL, K. (2020), «Wrongfully Accused by an Algorithm», *The New York Times*, 24 de junio.
- HOBBS, T. (1996), *Leviathan*, Oxford, Oxford University Press.
[Disponible en castellano como *Leviatán* (2003), Buenos Aires, Losada.]
- HOFFMAN, M. (2014), *Foucault and Power: The Influence of Political Engagement on Theories of Power*, Londres, Bloomsbury.
- HUGHES, J. (2004), *Citizen Cíborg: Why Democratic Societies Must Respond to the Redesigned Human of the Future*, Cambridge, MA, Westview Press.
- ILO (Organización Internacional del Trabajo) (2017), *Global Estimates of Modern Slavery*, Génova, Oficina de la Organización Internacional del Trabajo).
[Disponible en: <https://www.ilo.org/global/publications/books/WCMS_575479/lang-en/index.htm>; disponible en castellano como *Estimaciones mundiales sobre la esclavitud moderna* en: <https://www.ilo.org/wcmsp5/groups/public/@dgreports/@dcomm/documents/publication/wcms_651915.pdf>.]
- ISRAEL, T. (2020), *Facial Recognition at a Crossroads: Transformation at our Borders & Beyond*, Ottawa, Samuelson-Glushko Canadian Internet Policy & Public Interest Clinic. [Disponible en: <https://cippic.ca/uploads/FR_Transforming_Borders-OVERVIEW.pdf>.]
- JAVANBAKHT, A. (2020), «The Matrix Is Already There: Social Media Promised to Connect Us, But Left Us Isolated, Scared, and Tribal», *The Conversation*, 12 de noviembre. [Disponible en: <<https://theconversation.com/the-matrix-is-already-here-social-media-promised-to-connect-us-but-left-us-isolated-scared-and-tribal-148799>>.]
- JONAS, H. (1984), *The Imperative of Responsibility: In Search of an Ethics for the Technological Age*, Chicago, University of Chicago Press.

- [Disponible en castellano como *El principio de responsabilidad: ensayo de una ética para la civilización tecnológica* (1995), Barcelona, Herder.]
- KAFKA, F. (2009), *The Trial*, trad. por M. Mitchell, Oxford, Oxford University Press. [Disponible en castellano como *El proceso* (2016), Madrid, Valdemar.]
- KARPPI, T.; KÄHKÖNEN, L.; MANNEVUO, M.; PAJALA, M. y SIHVONEN, T. (2016), «Affective Capitalism: Investments and Investigations», *Ephemera: Theory & Politics in Organization* 16(4), págs. 1-13.
- KENNEDY, H.; STEEDMAN, R. y JONES, R. (2020), «Approaching Public Perceptions of Datafication through the Lens of Inequality: A Case Study in Public Service Media», *Information, Communication & Society*. [Disponible en: <<https://doi.org/10.1080/1369118X.2020.1736122>>.]
- KINKEAD, D. y DOUGLAS, D. M. (2020), «The Network and the Demos: Big Data and the Epistemic Justifications of Democracy», en K. McNish y J. Gailliot (eds.), *Big Data and Democracy*, Edimburgo, Edinburgh University Press, págs. 119-133.
- KLEEMAN, S. (2015), «Woman Charged with False Reporting after Her Fitbit Contradicted Her Rape Claim», *Mic.com*, 25 de junio. [Disponible en: <<https://www.mic.com/articles/121319/fitbit-rape-claim>>.]
- KORINEK, A. y STIGLITZ, J. (2019), «Artificial Intelligence and Its Implications for Income Distribution and Unemployment», en A. Agrawal, J. Gans y A. Goldfarb (eds.), *The Economics of Artificial Intelligence: An Agenda*, Chicago, University of Chicago Press, págs. 349-390.
- KOZEL, S. (2007), *Closer: Performance, Technologies, Phenomenology*, Cambridge, MA, MIT Press.
- KURZWEIL, R. (2005), *The Singularity Is Near: When Humans Transcend Biology*, Nueva York, Viking. [Disponible en castellano como *La singularidad está cerca* (2021), Berlín, Lola books.]
- KWET, M. (2019), «Digital Colonialism Is Threatening the Global South», *Al Jazeera*, 13 de marzo. [Disponible en: <<https://www.aljazeera.com/indepth/opinion/digital-colonialism-threatening-global-south-190129140828809.html>>.]
- LACLAU, E. (2005), *On Populist Reason*, Nueva York, Verso. [Disponible en castellano como *La razón populista* (2005), Madrid, Fondo de Cultura

Económica de España.]

- LAGERKVIST, A. (ed.) (2019), *Digital Existence: Ontology, Ethics and Transcendence in Digital Culture*, Abingdon, Routledge.
- LANIER, J. (2010), *You Are Not a Gadget: A Manifesto*, Nueva York, Borzoi Books.
- LARSON, J.; MATTU, S.; KIRCHNER, L. y ANGWIN, J. (2016), «How We Analyzed the COMPAS Recidivism Algorithm», *ProPublica*, 23 de mayo. [Disponible en: <<https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm>>.]
- LASH, S. (2007), «Power after Hegemony», *Theory, Culture & Society* 24(3), págs. 55-78.
- LATOUR, B. (1993), *We Have Never Been Modern*, trad. por C. Porter, Cambridge, MA, Harvard University Press. [Disponible en castellano como *Nunca fuimos modernos* (2007), Buenos Aires, Siglo XXI.]
- (2004), *Politics of Nature: How to Bring the Sciences into Democracy*, trad. por C. Porter, Cambridge, MA, Harvard University Press. [Disponible en castellano como *Políticas de la naturaleza* (2013), Barcelona, RBA.]
- LAZZARATO, M. (1996), «Immaterial Labor», en P. Virno y M. Hardt (eds.), *Radical Thought in Italy: A Potential Politics*, Mineápolis, University of Minnesota Press, págs. 142-157.
- (2014), *Signs and Machines: Capitalism and the Production of Subjectivity*, trad. por J. D. Jordan, Los Angeles, Semiotext(e). [Disponible en castellano como *Signos y máquinas: el capitalismo y la producción de la subjetividad* (2020), Madrid, Enclave de libros.]
- LEOPOLD, A. (1949), *A Sand County Almanac*, Nueva York, Oxford University Press.
- LIAO, S. M. (ed.) (2020), *Ethics of Artificial Intelligence*, Nueva York, Oxford University Press.
- LIN, P.; ABNEY, K. y JENKINS, R. (eds.) (2017), *Robot Ethics 2.0*, Nueva York, Oxford University Press.
- LLANSÓ, E. J. (2020), «No Amount of “AI” in Content Moderation Will Solve Filtering’s Prior-Restraint Problem», *Big Data & Society* 7(1). [Disponible en: <<https://doi.org/10.1177/2053951720920686>>.]
- LOIZIDOU, E. (2007), *Judith Butler: Ethics, Law, Politics*, Nueva York, Routledge.

- LUKES, S. (2019), «Power, Truth and Politics», *Journal of Social Philosophy* 50(4), págs. 562-576.
- LYON, D. (1994), *The Electronic Eye*, Mineápolis, University of Minnesota Press. [Disponible en castellano como *El ojo electrónico: el auge de la sociedad de la vigilancia* (1995), Madrid, Alianza.]
- (2014), «Surveillance, Snowden, and Big Data: Capacities, Consequences, Critique», *Big Data & Society* 1(2). [Disponible en: <<https://doi.org/10.1177/2053951714541861>>.]
- MACKENZIE, A. (2002), *Transductions: Bodies and Machines at Speed*, Londres, Continuum.
- MACKINNON, R.; HICKOK, E.; BAR, A. y LIM, H. (2014), «Fostering Freedom Online: The Role of Internet Intermediaries», París, Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura (UNESCO). [Disponible en: <<http://www.unesco.org/new/en/communication-and-information/resources/publications-and-communication-materials/publications/full-list/fostering-freedom-online-the-role-of-internet-intermediaries/>>.]
- MAGNANI, L. (2013), «Abducting Personal Data, Destroying Privacy», en M. Hildebrandt y K. de Vries (eds.), *Privacy, Due Process, and the Computational Turn*, Nueva York, Routledge, págs. 67-91.
- MANN, S.; NOLAN, J. y WELLMAN, B. (2002), «Sousveillance: Inventing and Using Wearable Computing Devices for Data Collection in Surveillance Environments», *Surveillance & Society* 1(3), págs. 331-355.
- MARCUSE, H. (2002), *One-Dimensional Man: Studies in the Ideology of Advanced Industrial Society*, Londres, Routledge. [Disponible en castellano como *El hombre unidimensional* (2016), Barcelona, Austral.]
- MARTÍNEZ-BASCUÑÁN, M. (2016), «Misgivings on Deliberative Democracy: Revisiting the Deliberative Framework», *World Political Science* 12(2), págs. 195-218.
- MARX, K. (1977), *Economic and Philosophic Manuscripts of 1844*, trad. por M. Milligan, Moscú, Progress Publishers. [Disponible en castellano como *Manuscritos de economía y filosofía* (2013), Madrid, Alianza.]
- (1990), *Capital: A Critique of Political Economy. Vol. 1*, trad. por B. Fowkes, Londres, Penguin. [Disponible en castellano como *El capital*

- (tomo 1/vol. 1): *crítica de la economía política* (2007), México DF, Siglo XXI.]
- MASSUMI, B. (2014), *What Animals Teach Us about Politics*, Durham, NC, Duke University Press.
- MATZNER, T. (2019), «Plural, Situated Subjects in the Critique of Artificial Intelligence», en A. Sudmann (ed.), *The Democratization of Artificial Intelligence: Net Politics in the Era of Learning Algorithms*, Bielefeld, Transcripción, págs. 109-122.
- MCCARTHY-JONES, S. (2020), «Artificial Intelligence Is a Totalitarian's Dream -Here's How to Take Power Back», *Global Policy*, 13 de agosto. [Disponible en: <https://www.globalpolicyjournal.com/blog/13/08/2020/artificial-intelligence-totalitarians-dream-heres-how-take-power-back>.]
- MCDONALD, H. P. (2003), «Environmental Ethics and Intrinsic Value», en H. P. McDonald (ed.), *John Dewey and Environmental Philosophy*, Albany, SUNY Press, págs. 1-56.
- MCKENZIE, J. (2001), *Perform or Else: From Discipline to Performance*, Nueva York, Routledge.
- MCNAY, L. (2008), *Against Recognition*, Cambridge, Polity.
- (2010), «Feminism and Post-Identity Politics: The Problem of Agency», *Constellations* 17(4), págs. 512-525.
- MCQUILLAN, D. (2019), «The Political Affinities of AI», en A. Sudmann (ed.), *The Democratization of Artificial Intelligence: Net Politics in the Era of Learning Algorithms*, Bielefeld, Transcript, págs. 163-173.
- MCSTAY, A. (2018), *Emotional AI: The Rise of Empathic Media*, Londres, Sage Publications.
- MIESSEN, M. y RITTS, Z. (eds.) (2019), *Para-Platforms: On the Spatial Politics of Right-Wing Populism*, Berlín, Sternberg Press.
- MILL, J. S. (1963), *The Subjection of Women*, en J. M. Robson (ed.), *Collected Works of John Stuart Mill*, Toronto, Routledge. [Disponible en castellano como *La esclavitud femenina* (2008), La Laguna, Artemisa.]
- (1978), *On Liberty*, Indianápolis, Hackett Publishing. [Disponible en castellano como *Sobre la libertad* (2008), Madrid, Tecnos.]
- MILLER, D. (2003), *Political Philosophy: A Very Short Introduction*, Oxford, Oxford University Press. [Disponible en castellano como *Filosofía política: una breve introducción* (2011), Madrid, Alianza.]

- MILLS, C. W. (1956), *The Power Elite*, Nueva York, Oxford University Press. [Disponible en castellano como *La elite del poder* (2013), México, Fondo de Cultura Económica.]
- MOFFITT, B. (2016), *Global Rise of Populism: Performance, Political Style, and Representation*, Stanford, Stanford University Press.
- MOORE, J. W. (2015), *Capitalism in the Web of Life: Ecology and the Accumulation of Capital*, Londres, Verso. [Disponible en castellano como *El capitalismo en la trama de la vida* (2020), Madrid, Traficante de sueños.]
- MOORE, P. (2018), *The Quantified Self in Precarity: Work, Technology and What Counts*, Nueva York, Routledge.
- MORAVEC, H. (1988), *Mind Children: The Future of Robot and Human Intelligence*, Cambridge, MA, Harvard University Press.
- MOROZOV, E. (2013), *To Save Everything, Click Here: Technology, Solutionism, and the Urge to Fix Problems That Don't Exist*, Londres, Penguin.
- MOUFFE, C. (1993), *The Return of the Political*, Londres, Verso. [Disponible en castellano como *El retorno de lo político* (2021), Barcelona, Paidós Ibérica.]
- (2000), *The Democratic Paradox*, Londres, Verso. [Disponible en castellano como *La paradoja democrática* (2012), Barcelona, Gedisa.]
- (2005), *On the Political: Thinking in Action*, Londres, Routledge. [Disponible en castellano como *En torno a lo político* (2007), Buenos Aires, Fondo de Cultura Económica.]
- (2016), «Democratic Politics and Conflict: An Agonistic Approach», *Política Común*. [Disponible en: <<http://dx.doi.org/10.3998/pc.12322227.0009.011>>.]
- MURRAY, D. (2019), *The Madness of the Crowds: Gender, Race and Identity*, Londres, Bloomsbury.
- NÆSS, A. (1989), *Ecology, Community and Lifestyle: Outline of an Ecosophy*, editado y trad. por D. Rothenberg, Cambridge, Cambridge University Press. [Disponible en castellano como *Ecología, comunidad y estilo de vida* (2000), Buenos Aires, Prometeo Libros.]
- NEMITZ, P. F. (2018), «Constitutional Democracy and Technology in the Age of Artificial Intelligence», *SSRN Electronic Journal* 376(2133). [Disponible en: <<https://doi.org/10.2139/ssrn.3234336>>.]

- NGUYEN, C. T. (2020), «Echo Chambers and Epistemic Bubbles», *Episteme* 17(2), págs. 141-161.
- NIELSEN, K. (1989), «Marxism and Arguing for Justice», *Social Research* 56(3), págs. 713-739.
- NIYAZOV, S. (2019), «The Real AI Threat to Democracy», *Towards Data Science*, 15 de noviembre. [Disponible en: <<https://towardsdatascience.com/democracys-unsettling-future-in-the-age-of-ai-c47b1096746e>>.]
- NOBLE, S. U. (2018), *Algorithms of Oppression: How Search Engines Reinforce Racism*, Nueva York, New York University Press.
- NOZICK, R. (1974), *Anarchy, State, and Utopia*, Nueva York, Basic Books. [Disponible en castellano como *Anarquía, Estado y utopía* (2015), Londres, Innisfree.]
- NUSSBAUM, M. (2000), *Women and Human Development: The Capabilities Approach*, Cambridge, Cambridge University Press. [Disponible en castellano como *Las mujeres y el desarrollo humano* (2012), Barcelona, Herder.]
- (2006), *Frontiers of Justice: Disability, Nationality, Species Membership*, Cambridge, MA, Harvard University Press. [Disponible en castellano como *Las fronteras de la justicia* (2012), Barcelona, Paidós Ibérica.]
- (2016), *Anger and Forgiveness: Resentment, Generosity, Justice*, Nueva York, Oxford University Press. [Disponible en castellano como *La ira y el perdón: resentimiento, generosidad, justicia* (2018), México, Fondo de Cultura Económico.]
- NYHOLM, S. (2020), *Humans and Robots: Ethics, Agency, and Anthropomorphism*, Londres, Rowman & Littlefield.
- O'NEIL, C. (2016), *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*, Nueva York, Crown Books. [Disponible en castellano como *Armas de destrucción matemática: cómo el big data aumenta la desigualdad y amenaza la democracia* (2018), Madrid, Capitán Swing.]
- OTT, K. (2020), «Grounding Claims for Environmental Justice in the Face of Natural Heterogeneities», *Erde* 151(2-3), págs. 90-103.
- OWE, A.; BAUM, S. D. y COECKELBERGH, M. (por publicar), «How to Handle Nonhumans in the Ethics of Artificial Entities: A Survey of the Intrinsic

Valuation of Nonhumans».

- PAPACHARISSI, Z. (2011), *A Networked Self: Identity, Community and Culture on Social Network Sites*, Nueva York, Routledge.
- (2015), *Affective Publics: Sentiment, Technology, and Politics*, Oxford, Oxford University Press.
- PARIKKA, J. (2010), *Insect Media: An Archaeology of Animals and Technology*, Mineápolis, University Of Minnesota Press.
- PARISER, E. (2011), *The Filter Bubble*, Londres, Viking. [Disponible en castellano como *El filtro burbuja* (2021), Barcelona, Taurus.]
- PARVIAINEN, J. (2010), «Choreographing Resistances: Kinaesthetic Intelligence and Bodily Knowledge as Political Tools in Activist Work», *Mobilities* 5(3), págs. 311-330.
- PARVIAINEN, J. y COECKELBERGH, M. (2020), «The Political Choreography of the Sophia Robot: Beyond Robot Rights and Citizenship to Political Performances for the Social Robotics Market», *AI & Society*. [Disponible en: <<https://doi.org/10.1007/s00146-020-01104-w>>.]
- PASQUALE, F. A. (2019), «Data-Informed Duties in AI Development» 119 *Columbia Law Review* 1917 (2019), U of Maryland Legal Studies Research Paper No. 2019-14. [Disponible en SSRN: <<https://ssrn.com/abstract=3503121>>.]
- PESSACH, D. y SHMUELI, E. (2020), «Algorithmic Fairness». [Disponible en: <https://arxiv.org/abs/2001.09784>.]
- PICARD, R. W. (1997), *Affective Computing*, Cambridge, MA, MIT Press.
- PIKETTY, T.; SAEZ, E. y STANTCHEVA, S. (2011), «Taxing the 1%: Why the Top Tax Rate Could Be over 80%», *VOXEU/CEPR*, 8 de diciembre. [Disponible en: <<https://voxeu.org/article/taxing-1-why-toptax-rate-could-be-over-80>>.]
- POLONSKI, V. (2017), «How Artificial Intelligence Conquered Democracy», *The Conversation*, 8 de agosto. [Disponible en: <<https://theconversation.com/how-artificial-intelligence-conquered-democracy-77675>>.]
- PUSCHMANN, C. (2018), «Beyond the Bubble: Assessing the Diversity of Political Search Results», *Digital Journalism* 7(6), págs. 824-843.
- RADAVOI, C. N. (2020), «The Impact of Artificial Intelligence on Freedom, Rationality, Rule of Law and Democracy: Should We Not Be Debating

- It?», *Texas Journal on Civil Liberties & Civil Rights* 25(2), págs. 107-129.
- RANCIÈRE, J. (1991), *The Ignorant Schoolmaster*, trad. por K. Ross. Stanford, Stanford University Press. [Disponible en castellano como *El maestro ignorante* (2010), Barcelona, Laertes.]
- (1999), *Disagreement*, trad. por J. Rose, Mineápolis, University of Minnesota Press.
- (2010), *Dissensus*, trad. por S. Corcoran. Nueva York, Continuum. [Disponible en castellano como *El desacuerdo* (1996), Buenos Aires, Nueva visión Argentina.]
- RAWLS, J. (1971), *A Theory of Justice*, Oxford, Oxford University Press. [Disponible en castellano como *Teoría de la justicia* (1999), Barcelona, Fondo de Cultura Económica.]
- (2001), *Justice as Fairness: A Restatement*, Cambridge, MA, Harvard University Press. [Disponible en castellano como *La justicia como equidad* (2012), Barcelona, Paidós Ibérica.]
- REGAN, T. (1983), *The Case for Animal Rights*, Berkeley, University of California Press. [Disponible en castellano como *En defensa de los derechos de los animales* (2016), Barcelona, Fondo de Cultura Económica.]
- RENSCH, A. T. L. (2019), «The White Working Class Is a Political Fiction», *The Outline*, 25 de noviembre. [Disponible en: <<https://theoutline.com/post/8303/white-working-class-political-fiction?zd=1&zi=oggsrqmd>>.]
- RHEE, J. (2018), *The Robotic Imaginary: The Human and the Price of Dehumanized Labor*, Minneapolis, University of Minnesota Press.
- RICOEUR, P. (1978), «The Metaphor Process as Cognition, Imagination, and Feeling», *Critical Inquiry* 5(1), págs. 143-159.
- RIEGER, S. (2019), «Reduction and Participation», en A. Sudmann (ed.), *The Democratization of Artificial Intelligence: Net Politics in the Era of Learning Algorithm*, Bielefeld, Transcripción, págs. 143-162.
- RIVERO, N. (2020), «The Pandemic is Automating Emergency Room Triage», *Quartz*, 21 de agosto. [Disponible en: <<https://qz.com/1894714/covid-19-is-boosting-the-use-of-ai-triage-in-emergency-rooms/>>.]

- RODEN, D. (2015), *Posthuman Life: Philosophy at the Edge of the Human*, Londres, Routledge.
- ROLNICK, D.; DONTI, P. L.; KAACK, L. H., *et al.* (2019), «Tackling Climate Change with Machine Learning». [Disponible en: <<https://arxiv.org/pdf/1906.05433.pdf>>.]
- ROLSTON, H. (1988), *Environmental Ethics: Duties to and Values in the Natural World*, Philadelphia, Temple University Press.
- RØNNOW-RASMUSSEN, T. y ZIMMERMAN, M. J. (eds.), (2005), *Recent Work on Intrinsic Value*, Dordrecht, Springer Netherlands.
- ROUSSEAU, J. J. (1997), *Of the Social Contract*, en V. Gourevitch (ed.), *The Social Contract and Other Later Political Writings*, Cambridge, Cambridge University Press, págs. 39-152. [Disponible en castellano como *El contrato social* (2004), Madrid, Istmo.]
- ROUVROY, A. (2013), «The End(s) of Critique: Data-Behaviourism vs. Due-Process», en M. Hildebrandt y K. de Vries (eds.), *Privacy, Due Process and the Computational Turn: The Philosophy of Law Meets the Philosophy of Technology*, Londres, Routledge, págs. 143-167.
- ROWLANDS, M. (2009), *Animal Rights: Moral Theory and Practice*, Basingstoke, Palgrave.
- SACO, D. (2002), *Cybering Democracy: Public Space and the Internet*, Minneapolis, University of Minnesota Press.
- SÆTRA, H. S. (2020), «A Shallow Defence of a Technocracy of Artificial Intelligence, Examining the Political Harms of Algorithmic Governance in the Domain of Government», *Technology in Society* 62. [Disponible en: <<https://doi.org/10.1016/j.techsoc.2020.101283>>.]
- SAMPSON, T. D. (2012), *Virality: Contagion Theory in the Age of Networks*, Mineápolis, University of Minnesota Press.
- SANDBERG, A. (2013), «Morphological Freedom-Why We Not Just Want It, but Need It», en M. More y M. Vita-More (eds.), *The Transhumanist Reader*, Malden, MA, John Wiley & Sons, págs. 56-64.
- SARTORI, G. (1987), *The Theory of Democracy Revisited*, Chatham, NJ: Chatham House Publishers. [Disponible en castellano como *Teoría de la democracia* (1988), Madrid, Alianza.]
- SATTAROV, F. (2019), *Power and Technology*, Londres, Rowman & Littlefield.

- SAURETTE, P. y GUNSTER, S. (2011), «Ears Wide Shut: Epistemological Populism, Argutainment and Canadian Conservative Talk Radio», *Canadian Journal of Political Science* 44(1), págs. 195-218.
- SCANLON, T. M. (1998), *What We Owe to Each Other*, Cambridge, MA, Harvard University Press. [Disponible en castellano como *Lo que nos debemos unos a otros: ¿qué significa ser moral?* (2003), Barcelona, Paidós Ibérica.]
- SEGEV, E. (2010), *Google and the Digital Divide: The Bias of Online Knowledge*, Oxford, Chandos.
- SHARKEY, A. y SHARKEY, N. (2012), «Granny and the Robots: Ethical issues in Robot Care for the Elderly», *Ethics and Information Technology* 14(1), págs. 27-40.
- SIMON, F. M. (2019), «“We Power Democracy”: Exploring the Promises of the Political Data Analytics Industry», *The Information Society* 35(3), págs. 158-169.
- SIMONITE, T. (2018), «When It Comes to Gorillas, Google Photos Remains Blind», *Wired*, 11 de enero. [Disponible en: <https://www.wired.com/story/when-it-comes-to-gorillas-google-photos-remains-blind/>.]
- SINGER, P. (2009), *Animal Liberation*, Nueva York, HarperCollins. [Disponible en castellano como *Liberación animal* (2011), Barcelona, Taurus.]
- SOLOVE, D. J. (2004), *The Digital Person: Technology and Privacy in the Information Age*, Nueva York, New York University Press.
- SPARROW, R. (2021), «Virtue and Vice in Our Relationships with Robots», *International Journal of Social Robotics* 13(1), págs. 23-29.
- STARK, L.; GREENE, D. y HOFFMANN, A. L. (2021), «Critical Perspectives on Governance Mechanisms for AI/ML Systems», en J. Roberge y M. Castell (eds.), *The Cultural Life of Machine Learning: An Incursion into Critical AI Studies*, Cham, Palgrave Macmillan, págs. 257-280.
- STIEGLER, B. (1998), *Technics and Time, 1: The Fault of Epimetheus*, trad. por R. Beardsworth y G. Collins, Stanford, Stanford University Press. [Disponible en castellano como *La técnica y el tiempo* (2003), Hondarribia, Hiru.]
- (2019), *The Age of Disruption: Technology and Madness in Computational Capitalism*, trad. por D. Ross. Cambridge, Polity.

- STILGOE, J., Owen, R. y Macnaghten, P. (2013), «Developing A Framework for Responsible Innovation», *Research Policy* 42(9), págs. 1568-1580.
- STRUBELL, E.; GANESH, A. y MCCALLUM, A. (2019), «Energy and Policy Considerations for Deep Learning in NLP». [Disponible en: <<https://arxiv.org/abs/1906.02243>>.]
- SUAREZ-VILLA, L. (2009), *Technocapitalism: A Critical Perspective on Technological Innovation and Corporatism*, Filadelfia, Temple University Press.
- SUDMANN, A. (ed.) (2019), *The Democratization of Artificial Intelligence: Net Politics in the Era of Learning Algorithms*, Bielefeld, Transcript.
- SUN, T.; GAUT, A.; TANG, S.; HUANG, Y.; EL SHEREIF, M.; ZHAO, J.; MIRZA, D.; BELDING, E.; CHANG, K. W. y WANG, W. Y. (2019), «Mitigating Gender Bias in Natural Language Processing: Literature Review», en A. Korhonen, D. Traum y L. Marquez (eds.), *Proceedings of the 57th Annual Meeting of the Association of Computational Linguistics*, págs. 1630-1640. [Disponible en: <<https://www.aclweb.org/anthology/P19-1159.pdf>>.]
- SUN, X.; WANG, N.; CHEN, C.; NI, J.-M.; AGRAWAL, A.; CUI, X.; VENKATARAMANI, S.; EL MAGHRAOUI, K. y SRINIVASAN, V. (2020), «Ultra-Low Precision 4-Bit Training of Deep Neural Networks», en H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan y H. Lin (eds.), *Advances in Neural Information Processing Systems 33 Pre-Proceedings*, Actas de la 34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada. [Disponible en: <<https://proceedings.neurips.cc/paper/2020/file/13b919438259814cd5be8cb45877d577-Paper.pdf>>.]
- SUNSTEIN, C. R. (2001), *Republic.com*, Princeton, Princeton University Press. [Disponible en castellano como *Republica.com* (2003), Barcelona, Paidós Ibérica.]
- SUSSER, D., ROESSLER, B. y NISSENBAUM, H. (2019), «Technology, Autonomy, and Manipulation», *Internet Policy Review* 8(2). [Disponible en: <<https://doi.org/10.14763/2019.2.1410>>.]
- SWIFT, A. (2019), *Political Philosophy*, Cambridge, Polity.
- TANGERMAN, V. (2019), «Amazon Used an AI to Automatically Fire Low-Productivity Workers», *Futurism*, 26 de abril. [Disponible en:

<<https://futurism.com/amazon-ai-fire-workers>>.]

THALER, R. H. y SUNSTEIN, C. R. (2009), *Nudge: Improving Decisions about Health, Wealth, and Happiness*. Revised edition, Londres, Penguin. [Disponible en castellano como *Un pequeño empujón (nudge): el impulso que necesitas para tomar las mejores decisiones en salud, dinero y felicidad* (2017), Barcelona, Taurus.]

THOMPSON, N.; HARARI, Y. N. y HARRIS, T. (2018), «When Tech Knows You Better Than You Know Yourself», *Wired*, 10 de abril. [Disponible en: <<https://www.wired.com/story/artificial-intelligence-yuval-noah-harari-tristan-harris/>>.]

THORSETH, M. (2008), «Reflective Judgement and Enlarged Thinking Online», *Ethics and Information Technology* 10, págs. 221-231.

TITLEY, G. (2020), *Is Free Speech Racist?*, Cambridge, Polity.

TOCQUEVILLE, A. (2000), *Democracy in America*, trad. por H. C. Mansfield y D. Winthrop, Chicago, University of Chicago Press. [Disponible en castellano como *La democracia en América* (2010), Madrid, Trotta.]

TOLBERT, C. J.; MCNEAL, R. S. y SMITH, D. A. (2003), «Enhancing Civic Engagement: The Effect of Direct Democracy on Political Participation and Knowledge», *State Politics and Policy Quarterly* 3(1), págs. 23-41.

TSCHAKERT, P. (2020), «More-Than-Human Solidarity and Multispecies Justice in the Climate Crisis», *Environmental Politics*. [Disponible en: <<https://doi.org/10.1080/09644016.2020.1853448>>.]

TUFEKCI, Z. (2018), «Youtube, the Great Radicalizer», *The New York Times*, 10 de marzo.

TURKLE, S. (2011), *Alone Together: Why We Expect More from Technology and Less from Each Other*, Nueva York, Basic Books.

UMBRELLO, S. y SORGNER, S. (2019), «Nonconscious Cognitive Suffering: Considering Suffering Risks of Embodied Artificial Intelligence», *Philosophies* 4(2). [Disponible en: <<https://doi.org/10.3390/philosophies4020024>>.]

ONU (Organización de las Naciones Unidas) (1948), Declaración Universal de Derechos Humanos. [Disponible en: <<https://www.un.org/en/about-us/universal-declaration-of-human-rights>>.]

ONU (Organización de las Naciones Unidas) (2018), «Promotion and Protection of the Right to Freedom of Opinion and Expression», Sesión

sesenta y tres, 29 de agosto. [Disponible en: <<https://www.undocs.org/A/73/348>>.]

UNICRI (Instituto Internacional de las Naciones Unidas para Investigaciones sobre la Delincuencia y la Justicia) e INTERPOL (Organización Internacional de Policía Criminal) (2019), *Artificial Intelligence and Robotics for Law Enforcement*, Turín y Lyon, UNICRI y INTERPOL. [Disponible en: <<https://www.europarl.europa.eu/cmsdata/196207/UNICRI%20-%20Artificial%20intelligence%20and%20robotics%20for%20law%20enforcement.pdf>>.]

VALLOR, S. (2016), *Technology and the Virtues*, Nueva York, Oxford University Press.

VAN DEN HOVEN, J. (2013), «Value Sensitive Design and Responsible Innovation», en R. Owen, J. Bessant y M. Heintz (eds.), *Responsible Innovation: Managing the Responsible Emergence of Science and Innovation in Society*, Londres, Wiley, págs. 75-83.

VAN DIJK, J. (2020), *The Network Society. Fourth edition*, Londres, Sage Publications.

VAN PARIJS, P. (1995), *Real Freedom for All*, Oxford, Clarendon Press. [Disponible en castellano como *Libertad real para todos* (1996), Barcelona, Paidós Ibérica].

VARELA, F.; THOMPSON, E. T. y ROSCH, E. (1991), *The Embodied Mind: Cognitive Science and Human Experience*, Cambridge, MA, MIT Press.

VÉLIZ, C. (2020), *Privacy Is Power: Why and How You Should Take Back Control of Your Data*, Londres, Bantam Press. [Disponible en castellano como *Privacidad es poder: datos, vigilancia y libertad en la era digital* (2021), Barcelona, Debate.]

VERBEEK, P. P. (2005), *What Things Do: Philosophical Reflections on Technology, Agency, and Design*, University Park, Pennsylvania State University Press.

VIDAL, J. (2011), «Bolivia Enshrines Natural World's Rights with Equal Status for Mother Earth», *The Guardian*, 10 de abril. [Disponible en: <<https://www.theguardian.com/environment/2011/apr/10/bolivia-enshrinesnatural-worlds-rights>>.]

VON SCHOMBERG, R. (ed.) (2011), *Towards Responsible Research and Innovation in the Information and Communication Technologies and*

- Security Technologies Fields*, Luxemburgo, Oficina de Publicaciones de la Unión Europea. [Disponible en: <<https://op.europa.eu/en/publication-detail/-/publication/60153e8a-0fe9-4911-a7f4-1b530967ef10>>.]
- WAHL-JORGENSEN, K. (2008), «Theory Review: On the Public Sphere, Deliberation, Journalism and Dignity», *Journalism Studies* 9(6), págs. 962-970.
- WALK FREE FOUNDATION (2018), *The Global Slavery Index*. [Disponible en: <<https://www.globalslaveryindex.org/resources/downloads/>>.]
- WALLACH, W. y ALLEN, C. (2009), *Moral Machines*, Nueva York, Oxford University Press.
- WARBURTON, N. (2009), *Free Speech: A Very Short Introduction*, Oxford, Oxford University Press.
- WEBB, A. (2019), *The Big Nine: How the Tech Titans and Their Thinking Machines Could Warp Humanity*, Nueva York, Hachette Book Group. [Disponible en castellano como *Los nueve gigantes: cómo las grandes tecnológicas amenazan al futuro de la humanidad* (2021), Barcelona, Península.]
- WEBB, M. (2020), *Coding Democracy: How Hackers Are Disrupting Power, Surveillance, and Authoritarianism*, Cambridge, MA, MIT Press.
- WESTLUND, A. (2009), «Rethinking Relational Autonomy», *Hypatia* 24(4), págs. 26-49.
- WINNER, L. (1980), «Do Artifacts Have Politics?», *Daedalus* 109(1), págs. 121-136.
- (1986), *The Whale and the Reactor*, Chicago, University of Chicago Press. [Disponible en castellano como *La ballena y el reactor* (2009), Barcelona, Gedisa.]
- WOLFE, C. (2010), *What Is Posthumanism?*, Mineápolis, University of Minnesota Press.
- (2013), *Before the Law: Humans and Other Animals in a Biopolitical Frame*, Chicago, University of Chicago Press.
- (2017), «Posthumanism Thinks the Political: A Genealogy of Foucault's The Birth of Biopolitics», *Journal of Posthuman Studies* 1(2), págs. 117-135.
- WOLFF, J. (2016), *An Introduction to Political Philosophy. Third edition*, Oxford, Oxford University Press. [Disponible en castellano como *Filosofía política: una introducción* (2001), Barcelona, Ariel.]

- YEUNG, K. (2016), «“Hypernudge”: Big Data as a Mode of Regulation by Design», *Information, Communication & Society* 20(1), págs. 118-136.
- YOUNG, I. (2000), *Inclusion and Democracy*, Oxford, Oxford University Press.
- ZIMMERMANN, A.; DI ROSA, E. y KIM, H. (2020), «Technology Can’t Fix Algorithmic Injustice», *Boston Review*, 9 de enero. [Disponible en: <<http://bostonreview.net/science-nature-politics/annette-zimmermann-elena-di-rosa-hochan-kim-technology-cant-fix-algorithmic>>.]
- ZOLKOS, M. (2018), «Life as a Political Problem: The Post-Human Turn in Political Theory», *Political Studies Review* 16(3), págs. 192-204.
- ZUBOFF, S. (2015), «Big Other: Surveillance Capitalism and the Prospects of an Information Civilization», *Journal of Information Technology* 30(1), págs. 75-89.
- (2019), *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*, Londres, Profile Books. [Disponible en castellano como *La era del capitalismo de la vigilancia* (2020), Barcelona, Paidós.]

⁴ Las fechas de última consulta de los hipervínculos que figuran a lo largo de este libro se corresponden con los meses de preparación de la edición original, publicada por Polity Press en febrero de 2022. (*N. del E.*)

Edición en formato digital: 2023

Ilustración de cubierta: Ana Coco

© This edition is published by arrangement with Polity Press Ltd, Cambridge

© De la traducción: Lucas Álvarez Canga, 2023

© Ediciones Cátedra (Grupo Anaya, S. A.), 2023

Calle Valentín Beato, 21

28037 Madrid

www.catedra.com

ISBN ebook: 978-84-376-4573-5

Está prohibida la reproducción total o parcial de este libro electrónico, su transmisión, su descarga, su descompilación, su tratamiento informático, su almacenamiento o introducción en cualquier sistema de repositorio y recuperación, en cualquier forma o por cualquier medio, ya sea electrónico, mecánico, conocido o por inventar, sin el permiso expreso escrito de los titulares del Copyright.